Predicting a successful loan in prosocial lending: the role of partner accreditation and storytelling on Kiva.org

Martina Pocchiari
Erasmus University Rotterdam

Master Thesis

Predicting a successful loan in prosocial lending: the role of partner accreditation and storytelling on Kiva.org

MSc Marketing Management
Rotterdam School of Management

June 19th, 2016

Coach: Pieter C. Schoonees
Co-reader: Alina Ferecatu

Student: Martina Pocchiari
Student nr.: 437283

Preface

The copyright of this master thesis rests with the author. The author is responsible for its contents. RSM is only responsible for the educational coaching and cannot be held liable for the content.

“And now I know that we must lift the sail
And catch the winds of destiny
Wherever they drive the boat.”
(E.L. Masters, 1916)

To my loving Family,
once more, and forever,
my only guiding light.

Executive Summary

Modern networking tools, such as the Internet or social media, have increased the ability of users around the world to connect with each other and facilitated the exchange of information and money. This trend has encouraged the emergence of online prosocial lending platforms and organizations, whose mission is to provide financial services to low income clients. Kiva.org, a non-profit organization, is one of them. Leveraging the Internet and a network of micro-finance institutions, Kiva connects people through lending, to alleviate poverty around the world. The prosocial mechanism, however, comes with a major drawback. Borrowers, who are especially based in developing countries, may have a hard time repaying their debts, and enter a so called "over-indebtness spiral". By asking loans to repay previous debt, borrowers' financial condition may become more and more critical. Consequences on their lives and on the parties to the transactions are dramatic. To reduce this threat, a model predicting the chances of loan repayment, given project-specific elements, can help screening potentially hazardous situations. The present research aims at providing the parties to the prosocial lending process a useful, yet simple, predictive model, that could individuate loan projects particularly exposed to a risk of default. In addition, this research expands and complements existing studies in the field, by adding two innovative predictors to the predictive models. For the first time since their introduction, the presence and number of Kiva's Social Performance Badges are tested as predictor to loan repayment. These badges give a measure of the social committment of Kiva's partner institutions, as well as a measure of social return on the project for the lenders. The second innovative predictive element is the emotional valence of the projects' text descriptions, to follow and complement the work of Dorfleitner et al. (2016) and Gao and Lin, (2015). Two general predictive models are developed at the beginning of the paper for practitioners' use (Model 0 and Model 1). The aim is to obtain the highest predictive accuracy while conserving interpretability. The model with least predictors (Model 1) is then suggested for practical use. To explore and discuss more in depth the impact of Social Performance Badges on chances of loan repayment, two additional models (Model A and B) are developed on two sets of loans, split by starting dates. This allows to compare prediction dynamics before and after the introduction of the badges. From the insights gained from the comparison, a final model (Model C) is developed, to account for relevant interaction effects, likely to have arisen after the introduction of the badges. From the results of all the models, it appears that Social Performance Badges have predictive potential. Taken individually, the presence of the Client Voice, Family and Community, Enterpreneurial Support and Facilitation of Savings badges has a positive impact on chances of repayment. In aggregate, an increase in the number of badges corresponds to an increase in probability of loan default. The introduction of the badges created a positive interaction effect between their number and several project sectors (among others, Manufacturing, Health and Food) and a negative interaction effect with partners' profitability and the sentiment score of text descriptions. The findings about emotional valence of text descriptions are in line with Dorfleitner et al. (2016) and Gao and Lin (2015), and confirm that positive text descriptions are associated with higher chances of loan repayment. At the end of the paper, the remaining loan- and partner-specific predictors are summarized according to whether they increase or decrease the chances of loan repayment. Although subject to several limitations, the models developed in this research achieve satisfying prediction accuracy and can be both a useful instrument for practitioners and an

interesting starting point for further research about project- and partner-specific characteristics and platform-specific accreditation systems.

Table of Contents

1. Introduction

The increasing spread and potential of modern networking tools, such as the internet and social media platforms, is allowing a growing number of people to make their contribution towards social good. One of the ways in which individuals can make an impact in such an interconnected reality is undertaking *prosocial lending*. Broadly speaking, prosocial lending can be seen as an initiative to empower and connect people, while creating opportunities around the world, by means of micro-loans with very low interest rates.

As far as reported results are concerned, microfinance has proven itself to be very successful in this action of opportunity creation. It is not uncommon to encounter reported microcredit repayment rates as high as 95 per cent. This figure is much higher than the rates in the traditional credit markets (Fenton, 2010). But reported figures are often misleading and excessively optimistic. A study published on the Microbanking bulletin (2003) found that only 66 of the 124 microfinance institutions inspected were financially sustainable, in spite of the high repayment rates. (Microfinance Information Exchange Market, 2003). When a loan defaults in microfinance settings, usually it will never be paid back, and is a financial loss to the lenders who contributed to it. It is thus of uttermost importance for the financial sustainability of institutions, as well as in the interest of the lenders, to be able to discriminate more efficiently between successful and unsuccessful projects. Several studies have been carried out over the last few years to profile a successful microfinance project against a defaulting one. The present research is carried out in the specific context of a popular online prosocial lending platform, Kiva (www.kiva.org). Particular attention is paid to the impact of a novel accreditation system (the so called "Social Performance Badges") and of borrowers' storytelling on the chances of success of a loan.

1.1. From Fragmented Individuals to Social Good: Prosocial Lending

Prosocial lending is a form of micro-financing driven by the intention to affect the welfare of others, disconnected from the decision maker (Galak, Small and Stephen, 2011). More broadly, micro-finance refers to the provision of financial services to low-income clients (including self-employed) and its definition often entails both financial and social intermediation. As such, it has been defined as "not simply banking, but a development tool" (Ledgerwood, 1998). Micro-financing became very popular in the 1970s and 1980s, when it arose as a response to doubts and research findings about state-delivery of subsidized credit to poor farmers. Since then, the micro-finance practice evolved, making use of new platforms for social project management, development, coordination and communication. In particular, the advent and spread of the internet allowed for the creation of a marketplace for the democratization of loaning money, while reducing the time and distance of the transactions (Coleman, 2007). The process reached its peak when micro-financing institutions turned into interactive platforms, so that individuals could finance projects from anywhere, at any time, provided that they had computer access.

1.2. Kiva.org: Organization, Network and Field Partners

One of the most widely used, crowdfunded platforms for micro-finance is Kiva.org. Kiva is a non-profit organization with a mission to connect people through lending to alleviate poverty. Founded in 2005, today the organization operates in 83 different countries, counting on 450 volunteers. The recent figures on the Kiva website highlight an impressive total amount lent through the platform of nearly $806 million, the presence of almost 2,5 million Kiva users and a notable 98.42% successful loan repayment rate. Leveraging the Internet and a worldwide network of micro-finance institutions, Kiva lets individuals lend as little as $25, at a zero-percent interest rate, to help create opportunity around the world. Given the prosocial nature of its operations, the spread of its reach and the magnitude of the total lent capital, Kiva counts on more than 300 so-called "Field Partners". Field Partners are micro-finance organizations around the world "responsible for screening borrowers, posting loan requests to Kiva, disbursing loans and collecting repayments, and otherwise administering Kiva loans" (as of January 19th, 2016). Kiva's Partners work at the local level, within the communities where loans are being used to make a difference. While most of the Field Partners are micro-finance institutions, the list also includes schools, NGOs and social enterprises. An extensive list of Kiva's Field Partners can be retrieved at http://www.kiva.org/partners.

1.2.1. Kiva's Social Perfomance Badges

Kiva introduced Social Performance badges in 2011. *Social Performance* is defined by the Social Performance Task Force as "the effective translation of an institution's mission into practice in line with accepted social values". Kiva grants Social Performance badges specific to the areas of social performance a Field Partner demonstrates commitment to. The badges are assigned to Kiva Field Partners during an initial due diligence process and are updated annually, making up a system of accreditation within the operations of the platform. These tokens give users an insight into the work of the Field Partners and represent heuristics to quickly identify a Field Partner that is supporting communities in a way that is more meaningful to the user. A screenshot from the "Partners" page on Kiva.org can be found in the Appendix. Together with general partner's information, the Figure in Appendix also shows examples of one or more Social Performance badges.

1.3. What Lenders See: Loans Characteristics and Metrics

In order to participate in one of Kiva's funding projects, users can visit www.kiva.org website. When browsing the online interface, potential lenders have access to an extensive amount of information. The interface provides a first glance to the open loans available for funding. Each available loan carries an array of useful information, such as:

- name, nationality and occupation of the borrower(s),
- the amount to be funded,
- the percentage of funds already provided,
- a short text description, that could be, on clicking, expanded in a new browser tab.

In addition, users can sort results by country, gender, sector, groups versus individuals, attributes and others (i.e. matched versus unmatched loans). Once users click on a given loan, they have access to further information about the loan and the related field partner. Some funding projects also include a video that serves as an additional source of information for the users. Videos are uploaded and embedded in the website via YouTube. A screenshot of Kiva.org website interface can be found in Appendix. The richness of information provided through the platform leaves users with a wide range of inputs for choosing a project. Galak, Small and Stephen (2011) attempted to summarize the steps undertaken by Kiva users considering a funding decision. According to their research, the lender on Kiva goes through two fundamental decision-making moments:

(i) the lender chooses the borrower, through a list of potential projects currently seeking funding;

(ii) the lender selects an amount to lend, ranging from $25 to $5,000, in $25 increments.

Once the loan project receives the amount requested by one or more funders, the borrowers must repay the loan. The repayment must follow predetermined repayment schedules, and lenders are made aware of the repayment terms beforehand. Eventually, the research highlights a possible third decision moment. Once the money is paid back by the borrower(s), the sum is returned to the original lender. At this point, the user (lender) can:

(iii) withdraw the funds;

(iv) lend again to other borrowers.

Option (iv) would send the lender back to the first step of the decision-making process. **Figure 1** sums up the process that brings the loan from the lender(s) to the borrower(s), from both a user's and a borrower's perspective.

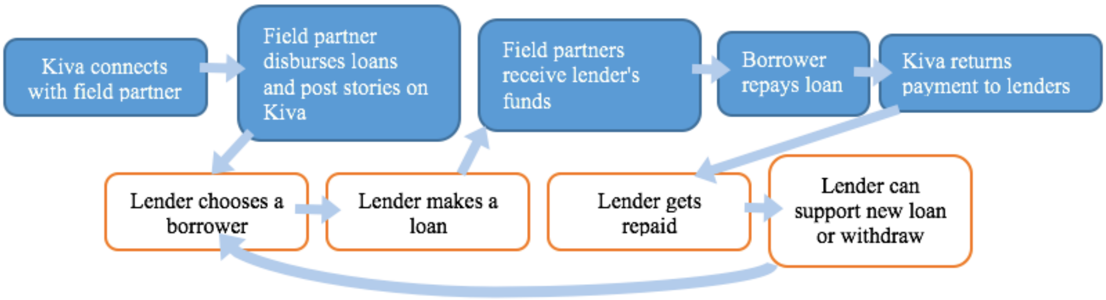

Kiva connects with field partner
Field partner disburses loans and post stories on Kiva
Field partners receive lender's funds
Borrower repays loan
Kiva returns payment to lenders
Lender chooses a borrower
Lender makes a loan
Lender gets repaid
Lender can support new loan or withdraw

FIGURE 1: THE PROSOCIAL LENDING PROCESS ON KIVA.ORG.

1.4. Loan Success or Default Metrics in Kiva User Interface

When consulting any project listed on Kiva, users can access information about the expected performance of the loan: this includes repayment term and schedule, possibility of currency exchange loss and data about the Field Partner in charge of the loan administration. Some general information about the country of origin of the borrower is also available, although not directly related to the expected performance. Field Partner information is particularly detailed and directly linked to the expected outcome of the project. Metrics include a 5-star scale risk rating, profitability measures, delinquency and default rates, loan-at-risk rates and currency exchange loss rates. Additionally, the number of total loans disbursed and the number of borrowers of the Field Partner is reported.

1.5. (In)success Stories: How Loans Default on Kiva.org

On the Kiva website, it is possible to find some definitions and criteria based on which loans are classified as defaulted. In the context, default (non-repayment) is defined as: *"the time when Kiva determines that collection of funds from a borrower or partner is doubtful, or when the cumulative amount repaid as of a quarterly reconciliation is less than the amount expected as of 180 days prior"*. The definition can, thus, be split in two. The cases in which a loan is considered in default are:

I) *The collection of funds from a borrower is doubtful*. Prior to defaulting, it is common to try to reschedule delinquent loans, in order to accommodate the eventual repayment of the loan. However, in spite of the organization's efforts to be flexible, sometimes borrowers fail to return funds, and the project ends in default.

II) *The cumulative amount repaid, as of a quarterly reconciliation, is less than the amount expected as of 180 days prior.* If, on a quarterly check, the borrower has returned less than an amount that was due 180 days before, the loan is classified as defaulted.

Furthermore, there is a third case in which a loan is considered unsuccessful:

III) *Field Partners have the option to default loans at any time, should they determine that further collection of loan repayments from the borrower is unlikely.* This decision is based on the discretion and free will of the field partner.

When a loan is defaulted, all lenders who contributed are notified. At that point, the lenders can consider the amount outstanding as a personal loss. In case of default, Field Partners are not allowed to cover what is owed to the lenders. The definitions above also imply that, for a loan to be considered successful, it must be fully repaid. Loans in fundraising, in repayment or inactive/ expired could, at any time on a quarterly basis, fall in default, in spite of their previous status. This information is crucial for the specification of research expectations and the subsequent definition of research variables.

2. Research Expectations

2.1. Prediction of Prosocial Lending Loan Success: State of the Art

Recently, several online microfinance platforms chose to make a big amount of data available to the public. Data from these sources is updated frequently and comes in convenient formats for statistical analyses. For this reason, research about factors predicting probabilities of loan success in a prosocial lending setting increased considerably with respect to the past. Several of these studies are focused on the dynamics behind lenders' decision-making in these new, evolving online environments. For what concerns the predicted probabilities of loan success, findings from existing literature are illustrated in **Table 1**. Results are divided into categories according to the type of predictor and the platform of reference. A more extensive review of the findings can be found in Section 3 of the present research.

Research Focus: Probability of Loan Repayment and Influencing Factors				
Platform & Author(s)	**Loan-Specific Predictors**	**Borrower-Specific Predictors**	**Lender-Specific Predictors**	**Chances of Loan Repayment**
Prosper - Kumar (2007)	Loan amount, credit grade	Group membership	Presence of an account verification	Decreased
Prosper - Everett (2015)	Loan amount	Group membership with personal relationships, degree of desperation	-	Decreased
	Measures of credit score, endorsement	Age	-	Increased
Smava and Auxmoney - Dorfleitner et al. (2016)	-	Text elements describing divorce and separation	-	Decreased
Prosper - Gao and Lin, (2015)	Positive loan requests and loans with more objective information	-	-	Increased

TABLE 1: FACTORS INFLUENCING LOAN SUCCESS IN PROSOCIAL LENDING

As Table 1 shows, elements influencing probabilities of loan success or default can refer either to the loan, to the borrowers advancing the financing request or to the lenders that take part in the lending project. From the research findings listed above, the increase in the amount requested by the borrowers appears to be a recurring predictor of loan default (Kumar, 2007 and Everett, 2015). Membership to a borrowing group has also been subject to prior research in Kumar (2007) and Everett (2015). In both cases, the feature was associated to projects more likely to default. The only predictors of successful loan repayment are the age of the borrower, a credit score received by the credit bureau and presence of personal endorsements. An increase is these variables is associated with less likelihood of default.

The variables used in the cited papers are platform-specific measures of project performance, as well as some account information. Examples of loan data from Prosper.com can be found at the link https://developers.prosper.com/docs/investor/loans-api/. Since loan, borrower and lender objects are highly specific to the policies and dynamics of the hosting platform, the present research aims at expanding the state of the art in predicting loan success, using data objects from the prosocial lending setting of Kiva.org.

2.2. Beyond the State of the Art: Social Performance Measures and Text Descriptions

In the present research, insights of previous literature about probabilities of loan success (Table 1) are used to derive a model which takes into account a large number of independent variables to predict probabilities of success or default of listed projects. An addition to the contribution of previous researches is the new, platform-specific presence of Social Performance Badges as factor to influence loan success. The assessment of the impact of Social Performance Badges takes into account the role of certifications and quality assessment for MFIs and applies it to a real-life, project-specific online setting. Moreover, the findings from Gao and Lin (2015) about the influence of positively-valenced text features on probability of loan defaults are taken as a reference point to assess the role of text descriptions on Kiva. The descriptions will be specific to projects listed on Kiva, which, unlike Prosper, is a non market-based player (Bruett, 2007).

Based on these elements, I expect that:

(a) among all the factors that may influence the loan success, the presence of at least one accreditation badge has a non-zero impact on the probabilities of loan repayment;

(b) the presence of emotion-loaded words in a text description positively influences loan success. In particular, I expect that positive text descriptions increase the chances of loan success, in line with the findings of Gao and Lin (2015) on Prosper.com.

2.3. Conceptual Model

To summarize the expectations entailed by the present research, a possible Conceptual Model is shown in **Figure 2.**

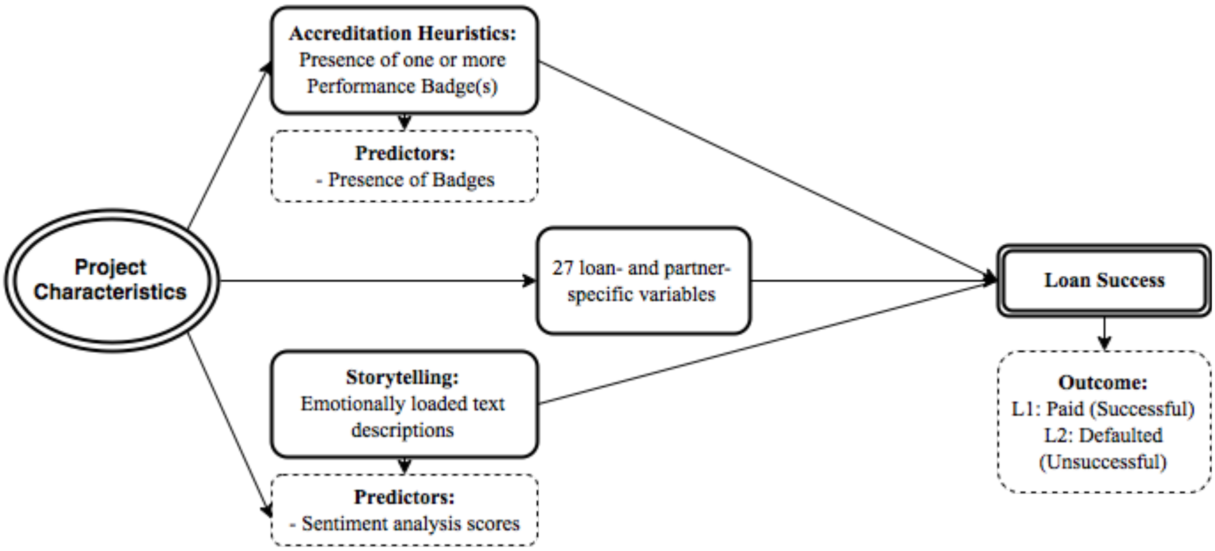

Accreditation Heuristics:
Presence of one or more Performance Badge(s)
Predictors:
- Presence of Badges
Project Characteristics
27 loan- and partner-specific variables
Loan Success
Storytelling:
Emotionally loaded text descriptions
Predictors:
- Sentiment analysis scores
Outcome:
L1: Paid (Successful)
L2: Defaulted (Unsuccessful)

FIGURE 2: CONCEPTUAL MODEL FOR THE STUDY OF LOAN SUCCESS BASED ON ACCREDITATION HEURISTICS AND VALENCE OF TEXT DESCRIPTIONS

Among all the factors that could drive and influence the loan success on Kiva.org, the present research focuses on accreditation heuristics (in particular, the presence of an accreditation system) and on the emotional content of storytelling material. A more detailed review of the literature about success factors, accreditation systems and text descriptions as soft success factors can be found in Section 3.

2.4. Managerial Relevance

In the domain of finance and financial analysis, one of the main decisions financial institutions have to make is to decide whether or not to grant a loan to a customer (Martens et al., 2007). The same is also true for the micro-finance industry, which, in recent times, has been struggling with frequent episodes and problems of over-indebtness[1] of the borrowers. On the long run, over-indebtness threatens to endanger both the social impact on microfinance institutions and the industry stability (Schicks, 2010). It is thus of uttermost importance to be able to have an a-priori estimate of the likelihood of default of microfinance projects. Predictive models can have a positive role in preventing dangerous cases of over-indebtness as well as losses sustained by all parties to the financing transactions. The present research extends several findings from previous literature about predicting likelihood of default to the online environment of Kiva.org. The research is particularly relevant for assessing the role of accreditation in an online setting, since the accreditation system introduced on the Kiva platform is relatively recent. At the moment, no previous research was conducted with a focus on Social Performance accreditation systems. Should the present expectations lead to encouraging results, the accreditation system could be improved, upgraded and possibly extended to more real-life, online environments. Furthermore, the present study incorporates emotional storytelling as an active predictor of loan repayment, in a context other than a market-player. The findings may enable more non-market oriented microlending platforms to adopt and enhance marketing techniques. In turn, this could help them to raise lenders' attention to particular causes and contribute to the overall success of the campaigns.

2.4.1. Field Partners

The research can interest several parties of the prosocial lending process. First of all, the field partners can benefit from the findings about the presence of Social Performance Badges. The findings would provide Field Partners with additional information about the dynamics that regulate loan success and the impact of these tokens. The badges would then serve not only as measures of social return but also as factors participating in campaign success. Field Partners presenting badges associated with higher chances of loan repayment could leverage this presence as an additional asset.

[1] A microfinance customer is over-indebted if "he is continuously struggling to meet repayment deadlines and repeatedly has to make unduly high sacrifices to meet his loan obligations" (Schicks, 2010).

2.4.2. Lenders

The second category which may have interest in the present research is that of the lenders to the loans. Lenders could extract valuable insights for their processes of loan selection from the present research. In fact, lenders can improve loan selection by collecting private information about their borrower customers (Fama, 1985). It is evident that a defaulted loan is a dangerous ground for a borrower who may fall in the over-indebtness trap. But it is also a loss sustained by the lender. It is in his interest, thus, to acknowledge factors contributing to a lower average default rate for the investment portfolio. Doing this would, in turn, also reduce overall lending costs (Everett, 2015).

2.4.3. Prosocial Lending Platforms

Eventually, the study of text descriptions in the context of a non-market player can extend the benefits already brought by Gao and Lin (2015) and encourage different platforms, market- or other non-market players to improve efficiency, by leveraging linguistic features of their prosocial offerings. From a marketing point of view, the analysis of text descriptions is particularly relevant for practitioners, since a potentially successful linguistic content can be thoughtfully crafted to highlight those elements which best convey the already existing probabilities of success.

3. Literature Review

3.1. Loan Success in Prosocial Lending Settings: Influencing Factors

The factors influencing the probabilities of loan success or default have been frequently investigated by researchers. Several studies were carried out both from a traditional microfinancing perspective (i.e. MFIs operating offline) and from a prosocial lending point of view. Prosocial lending studies were largely based, among others, on data from Prosper.com. The factors which proved to have significant effects on chances of loan repayment were both loan-specific and borrower-specific. Findings highlight the recurring presence of loan size, repayment terms, borrowers' groups and demographic specifications of the borrowing groups as significant influencing factors. More in detail, Godquin (2004) analyzed results from a quasi-experimental survey carried out in Bangladesh in 1991–1992. The study found that the provision of non-financial services has a positive impact on repayment performance. The size of the loan and the age of the borrowing group, instead, were found to have negative impact on repayments. Another study by Kumar (2007), based on data from Prosper.com, found that the loan amount, the credit grade, the presence of an account verification and a borrower group membership increase probability of loan default. Borrowing group dynamics on Prosper.com were further investigated by Everett (2015), whose study found that, when a borrower is a member of a group that has personal relationships, loan default rate is significantly lowered. This effect was explained by assuming that "group leader rewards incentivize the group leader to monitor the loans more carefully", which would in turn lower probabilities of default. According to the same research, other factors whose increase is likely to increase chances of default are, again, the amount requested and the degree of desperation of the borrower. Factors that, instead, decrease chances of default are measures of credit score (received by a credit bureau), the borrowers' age and the presence of endorsement features. For what concerns the repayment terms, no significant effect of type and pace of repayment schedule is reflected on client delinquency or default (Field and Pande, 2008).

3.2. Accreditation Systems and Probability of Default

3.2.1. Seals and Certifications in Microfinance

The issue of process reliability and verification is an actual and recurrent concern for both the financial and the micro-finance worlds. This is especially true following the economic downturns of the late 2000s. The case of Micro-Finance Institutions (MFIs) is particularly interesting since, over time, the category developed a strong social bottom-line brand image (Ashta and Bumacov, 2011). Today, it is not unfrequent for MFIs to undergo processes of certification and quality assessments, which indicate process quality and reduce the asymmetric information problems (Ashta and Bumacov, 2011). Ashta and Bumacov (2011), however, argue that having a quality certification may not be vital for a MFI, since borrowers involved would probably take loans as long as the process is cheaper than going to the money-lender. This is assumed to hold regardless of the presence of certifications and seals. Little or no research was carried out on project-specific accreditations. It may be of interest, thus, to investigate more forms of project-specific tokens and their relationships with project success.

3.2.2. Accreditation Systems in Online Prosocial Environments

Few studies exist about the role of accreditation systems in online prosocial environments, and, in particular, in platforms similar to Kiva.org. The reason may be that Kiva is one of the only platforms - if not the only - to grant accreditation instruments to its Field Partners. But the issue of social accreditation is a real concern, especially considering the complex nature of the prosocial lending insitutions. As reported by Ashta and Bumacov (2011), the economic model of microfinance is supposed to be the connecting piece between finance and charity, meeting the two-fold objective of financial and social returns. In such a setting, the traditional credit ratings would only assess the quality of the financial side of the MFI. Given the blank space in the measurement of the social returns, the recent social ratings could, instead, assess the quality of the social impact. As Ashta and Bumacov (2011) continue, "donors and socially oriented investors are willing to provide subsidized or free capital, or grants, in exchange of social return, so these stakeholders should be the main users of social rating reports". Kiva provides heuristics for the social return of each projects by means of the newly introduced Social Performance Badges. So far, this is the only form of project-specific social accreditation on a prosocial lending platform. Nonetheless, there are examples of several initiatives taken by platforms similar to Kiva oriented towards the recognition of a social return. Lendwithcare.org developed eligibility criteria for prospective micro-finance institutions that wish to partner with it. Zidisha.org has a "trust and security" section in which a verification of the identity of the featured entrepreneurs is alleged, and in which it is pointed out how Zidisha is "a purely online service without local offices or loan officers". Notably, watsi.org uploads a detailed list of transactions in a spreadsheet called "Transparency document", to prove that 100% of donation funds healthcare. Besides this, medical partners do not receive any form of accreditation. The same goes for milaap.org and unitedprosperity.org, which give a list of partners but do not provide discriminant features that resemble accreditation signs. With this respect, the present research could complement existing literature about repayment prediction. By including Social Performance Badges in a predictive model of loan success or default, this research can help to understand whether a project-specific accreditation system, besides providing more valuable information to social-oriented investors and assessing social return, can also concur to a model of loan success.

3.3. Text Descriptions and Loans Success on Prosocial Lending Platforms

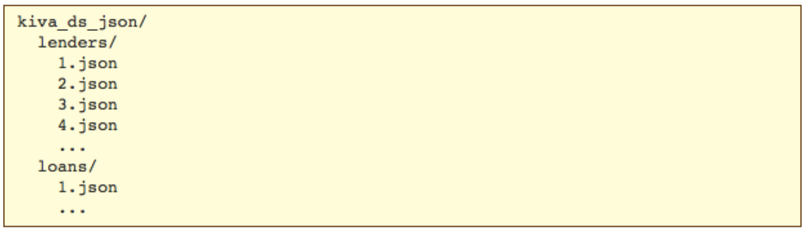
As reported by Lee and Lee (2012), lenders gain an emotional benefit from helping the poor by investing in their auctions, according to a survey conducted by Popfunding. In the same paper, further content analyses on prosocial lending platforms are encouraged to test this emotional impact. Studies performing content analysis on text descriptions accompaining loan projects are relatively recent, probably given that structured open source data have only been available since the past few years. In spite of the recency of the studies, researchers highlighted some valuable results that gave relevance to text descriptions and lexical elements as powerful predictive tools. Dorfleitner et al. (2016) investigated the role of text descriptions as soft factors for mitigating asymmetric information. The research is based on two European peer-to-peer lending platforms, Smava and Auxmoney. From the study, there is evidence that loan applicants using text elements describing divorce and separation, inherently negative, have a higher probability of default. Authors argue that the possible problems in the personal lives of the borrowers may actually affect their repayment behavior. However, the study also highlights how the relation of the description-text related soft factors to the default probability is much less strong than the relation of the same factors with the funding probability. A more recent research, based on data from Prosper.com, highights how positive loan requests and loans with more objective information are less likely to default (Gao and Lin, 2015), somewhat completing and reinforcing the findings from Dorfleitner et al. (2016). These recent findings have proven how text descriptions, taken as soft factors, can have valuable predictive power towards the probability of project success. The present research aims at confirming the conclusions above and, more particularly, at extending the findings from Gao and Lin (2015) to the setting of a prosocial lending, non market-based organization: Kiva.org.

4. Dataset

The data used for the present study can be retrieved from Kiva API at the web address http://build.kiva.org/docs/. Kiva's API is web-service based and resource-oriented. Data snapshots were downloaded from Kiva repository at the URL http://build.kiva.org/docs/data/snapshots (loans) and at http://api.kivaws.org/v1/partners.json (field partners) on January 1st, 2016.

4.1. Data Format

Data come in JSON or XML formats. Additionally, an HTML response format is allowed to access the data in common web browsers. In order to obtain the most out of the information, however, the JSON format is preferred for the retrieval of the data. Specialized data objects obtainable via Kiva API include information about loans, lenders, borrowers, loan updates, partners and lending teams. Simple data types include additional knowledge about dates, IDs, countries, locations and language codes. For the case of the loans, each data collection (snapshot) is a series of numbered files in JSON format. The number of each file in the collection corresponds to the page of data in the series (**Figure 3**).



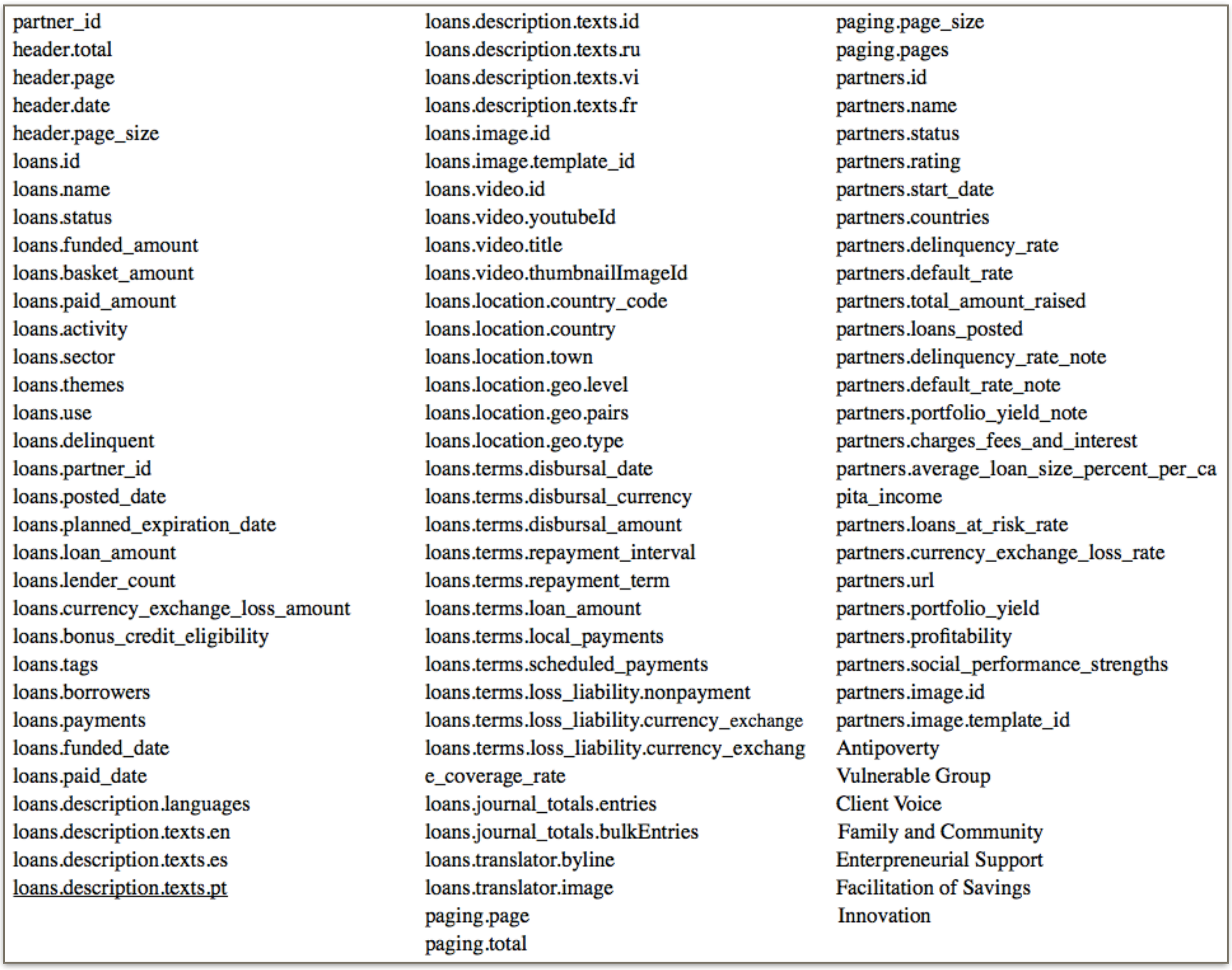
```
kiva_ds_json/
  lenders/
    1.json
    2.json
    3.json
    4.json
    ...
  loans/
    1.json
    ...
```

FIGURE 3: DATA SNAPSHOTS ARE DOWNLOADED AS A SERIES OF NUMBERED FILES IN JSON FORMAT

Data snapshots on loans (as of February 8th, 2016) are split into 2002 files. Field partners' data are, instead, stored in a single dataset. For the purposes of the present research, R Studio software is used to parse the data (R Core Team, 2015). All codes used are available the GitHub repository for this research project at the URL https://github.com/elisewinn/Pocchiari_2016_Thesis.

4.2. Starting Dataset

In order to have a general overview of the type and nature of the data stored in the snapshots, the first dataset (1.json) is analyzed in this section. This initial dataset, similarly to all the data from the snapshot, contains an extensive amount of information, gathered in 63 complex variables. **Figure 4** presents the initial column names and is helpful to broadly explore the variables stored in the snapshots. A complete description of the variables can be found in the Appendix.



partner_id
header.total
header.page
header.date
header.page_size
loans.id
loans.name
loans.status
loans.funded_amount
loans.basket_amount
loans.paid_amount
loans.activity
loans.sector
loans.themes
loans.use
loans.delinquent
loans.partner_id
loans.posted_date
loans.planned_expiration_date
loans.loan_amount
loans.lender_count
loans.currency_exchange_loss_amount
loans.bonus_credit_eligibility
loans.tags
loans.borrowers
loans.payments
loans.funded_date
loans.paid_date
loans.description.languages
loans.description.texts.en
loans.description.texts.es
loans.description.texts.pt
loans.description.texts.id
loans.description.texts.ru
loans.description.texts.vi
loans.description.texts.fr
loans.image.id
loans.image.template_id
loans.video.id
loans.video.youtubeId
loans.video.title
loans.video.thumbnailImageId
loans.location.country_code
loans.location.country
loans.location.town
loans.location.geo.level
loans.location.geo.pairs
loans.location.geo.type
loans.terms.disbursal_date
loans.terms.disbursal_currency
loans.terms.disbursal_amount
loans.terms.repayment_interval
loans.terms.repayment_term
loans.terms.loan_amount
loans.terms.local_payments
loans.terms.scheduled_payments
loans.terms.loss_liability.nonpayment
loans.terms.loss_liability.currency_exchange
loans.terms.loss_liability.currency_exchange_coverage_rate
loans.journal_totals.entries
loans.journal_totals.bulkEntries
loans.translator.byline
loans.translator.image
paging.page
paging.total
paging.page_size
paging.pages
partners.id
partners.name
partners.status
partners.rating
partners.start_date
partners.countries
partners.delinquency_rate
partners.default_rate
partners.total_amount_raised
partners.loans_posted
partners.delinquency_rate_note
partners.default_rate_note
partners.portfolio_yield_note
partners.charges_fees_and_interest
partners.average_loan_size_percent_per_capita_income
partners.loans_at_risk_rate
partners.currency_exchange_loss_rate
partners.url
partners.portfolio_yield
partners.profitability
partners.social_performance_strengths
partners.image.id
partners.image.template_id
Antipoverty
Vulnerable Group
Client Voice
Family and Community
Enterpreneurial Support
Facilitation of Savings
Innovation

FIGURE 4: NAMES OF THE VARIABLES AVAILABLE IN THE DATA SNAPSHOTS ON KIVA.ORG

Some of the variables, such as “loans.name” or “partners.countries”, are complex variables. This means that they are composed by nested lists or dataframes. For this reason, the actual number of columns after the unnesting procedures would be higher than the initial 63.

4.3. Missing Values

Before proceeding with any analysis on a larger dataset, missing values are assessed using a small sample of snapshots. The first two datasets (1.json and 2.json), two middle datasets (1000.json and 1100.json) and the last two datasets (2002.json and 2001.json) are analyzed. Each dataset contains 500 observations and a number of variables between 56 and 63, for a total number of 3000 observations. The choice to analyze different datasets is made to ensure that no systematic missing values are observed based only on differences in time periods, given that datasets are stored in ascending chronological order. As **Table 2** shows, a number of variables in the sample selected presents some recurring missing values. The table lists them in both absolute terms and as percentage of total observations in the sample.

Variable	Missing Values/ NAs	Percentage of Total* (%)
loans.basket_amount	2256	75,20%
loans.paid_amount	841	28,03%
loans.delinquent	2611	87,03%
loans.posted_date	42	1,40%
loans.currency_exchange_loss_amount	2711	90,37%
loans.funded_date	545	18,17%
loans.paid_date	1019	33,97%
loans.location.town	339	11,30%
loans.terms.loss_liability.currency_exchange_coverage_rate	799	26,63%
loans.video/thumbnail/youtube ID	2776	92,53%
partners.url	499	16,63%
partners.portfolio_yield	307	10,23%
partners.profitability	356	11,87%
	*N = 3000	

TABLE 2: ANALYSIS OF MISSING VALUES ON SELECTED SAMPLES

From this analysis, it is evident that some of the missing values are repeated across the sample. The variable “loans.basket_amount” is empty in half of the analyzed datasets and mostly empty in the other half. Some other missing values are time-dependent or dependent on the status of the loan. For example, the “loans.paid_amount” may present missing values if the loan is still in the funding phase. It is, in fact, impossible for the project owner to pay back money that haven’t been funded yet. This is particularly true in the last two datasets (2001 and 2002): since they are the most recent, they contain relatively new projects that haven’t yet completed the funding process. The variables presenting a problematic amount of missing values must be treated accordingly. The criterion to dispose of them is the extent to which they can contribute to the model of this research. Since the research aims at understanding which factors impact the success of a loan on Kiva.org, the problematic variables unlikely to explain loan success variance are excluded. More specifically, these would be variables that are not visible or are not directly influencing the repayment dynamics and processes. For the case of this research, the following variables are excluded from the analysis:

- **loans.basket_amount**: this is an amount that lenders have saved for potential later purchases but has not been disbursed yet. It does not contribute to the success of a project because the amount is not disbursed at all. It is rather an information that the lender has about its own virtual wallet and disposable income.
- **loans.paid_amount**: the variable is not present in all datasets, adding complexity to the analysis.

- **loans.delinquent**: removed due to an excessive number of missing values.
- **loans.currency_exchange_loss_amount**: this is a loss to the lenders and depends on economic fluctuations in the currency market. It also presents an excessive amount of missing values to be considered explanatory.
- **loans.funded_date**: missing values correspond to projects that have not yet been funded, and that are, hence, excluded from the scope of the research.
- **loans.paid_date**: this variable correlates perfectly with the dependent variable.
- **loans.location_town**: this information is very specific. Relevant information is likely to be contained in loans.country - which largely presents no missing values.
- **loans.terms.loss_liability.currency_exchange_coverage_rate**: this variable is not present in all datasets and complicates the binding process necessary for subsequent analyses.
- **loans.video** (or related variables): nearly 100% missing values, the variable is excluded because non explanatory.

The other variables listed in Table 2 and not mentioned for removal present a relatively low percentage of missing values that can be treated on an individual basis. This step is performed once the final dataset is assembled, after the cleaning phase. All the variables not shown either in Table 2 or in the bulleted list above present less than 10 missing values. This amount is assumed to be not problematic and the missing values are removed case-wise.

4.4. Dataset Cleaning and Appending Phases

As previously mentioned, the data snapshots are originally divided according to paging criteria and splitted in 2002 files of approximately 500 observations each. For the purposes of this research, it is ideal to append several of these files, to create a more extensive dataset with a higher number of observations. This makes it possible to include more observations and run analyses only once, to make the process more explanatory and time-efficient. To achieve this objective, 100 loans' datasets are appended (1-21.json, 23-50.json and 1953-2002.json). This collection of 100 datasets is then merged with the partners' dataset (partners.json) by partner_id. 50 oldest loan datasets are chosen along with 50 newest, to allow for comparison of successful projects over time, avoid seasonality problems and to account for the introduction of new decision-influencing elements — such as the Social Perfomance Badges. The datasets contain projects from 2006 to 2015. Since the 100 datasets present a non-fixed number of columns, they must be cleaned before the appending procedure. The cleaning phase does not only even the number of columns, but also discard those variables that are likely to be redundant, non-explanatory or noisy. **Table 3** presents the list of variables that are dropped from each of the 100 datasets and a brief rationale for the choice.

Variable Category	Single Variables Dropped	Rationales
Headers and Paging	"header.total","header.page", "header.page_size", "header.date", "paging.page", "paging.total", "paging.page_size", "paging.pages"	Data referring to paging and not to lending processes.
Ids	"loans.partner_id"	Redundant (partner_id is a sorting variable).
Text Descriptions	"loans.description.languages", "loans.description.texts.es/fr/ru/pt/id/vi/ar"	Text descriptions in foreign languages are not studied in this research.
Media	"loans.image.id", "loans.image.template_id"	The research is not concerned with the presence of multimedia elements in the projects, and the variables have too many NAs.
Activity and Use	"loans.activity", "loans.use", "loans.themes"	Correlated with "sector" and more complex.
Locations	"loans.location.geo.level", "loans.location.geo.type", "loans.location.country_code", "location.geo.pairs"	Largely correlated to "country" or too specific.
Other Information	"loans.planned_expiration_date", "loans.tags" “loans.name”, “borrower_name”	Present too many NAs. "loans.tags" stored as list(). "loans.name" substituted with unlisted column “names”. “borrower_name” has too many level to be effectively studied.
Terms	"loans.terms.disbursal_date", "loans.terms.disbursal_currency", "loans.terms.disbursal_amount", "loans.terms.loan_amount"*,"loans.terms.local_payments", “loans.terms.scheduled_payments"*, “loans.payments", "loans.terms.loss_liability.currency_exchange"	Either correlated to other variables (*) or introduced in different moments and not applicable to all datasets.
Translators	"loans.translator.byline", "loans.translator.image", "loans.journal_totals.bulkEntries"	Only using texts in English.
Partners	"partners.name", "partners.delinquency_rate_note"*, "partners.default_rate_note"*, "partners.default_rate_note"*, "partners.portfolio_yield_note"*, "partners.url"**, "partners.image.id"**, "partners.image.template_id"**, "partners.currency_exchange_loss_rate"^, "partners.social_performance_strengths"**	Either largely empty (*), or unexplicative/redundant for the analysis (**) or non-present in all the relevant datasets (^)

TABLE 3: VARIABLES EXCLUDED FROM THE ANALYSES AND RATIONALES

As Table 3 shows, many variables were dropped from the datasets because related to paging, ids, translations and, more in general, to factors not directly impacting the success of a loan. After removing these columns, all the 100 datasets end up containing the same number of variables and this makes it possible for the concatenating process to take place.

4.5. Remaining Missing Values and Correction of Measurement Levels

Even after the cleaning procedures, it appears that some variables present missing values **(Table 4)**.

Variable	NMiss	N	Proportion	Treatment
loans.terms.repayment_interval	77	41488	0,19%	Replace with mode
loans.terms.repayment_term	84	41488	0,20%	Replace with mean
partners.delinquency_rate	357	41488	0,86%	Replace with mean
partners.default_rate	357	41488	0,86%	Replace with mean
loans.description.texts.en	1072	41488	2,58%	Drop variable
partners.rating	1614	41488	3,89%	Replace with mean
partners.portfolio_yield	6318	41488	12,38%	Replace with mean
partners.profitability	6363	41488	15,23%	Replace with mean

TABLE 4: ANALYSIS AND DISPOSAL OF REMAINING MISSING VALUES

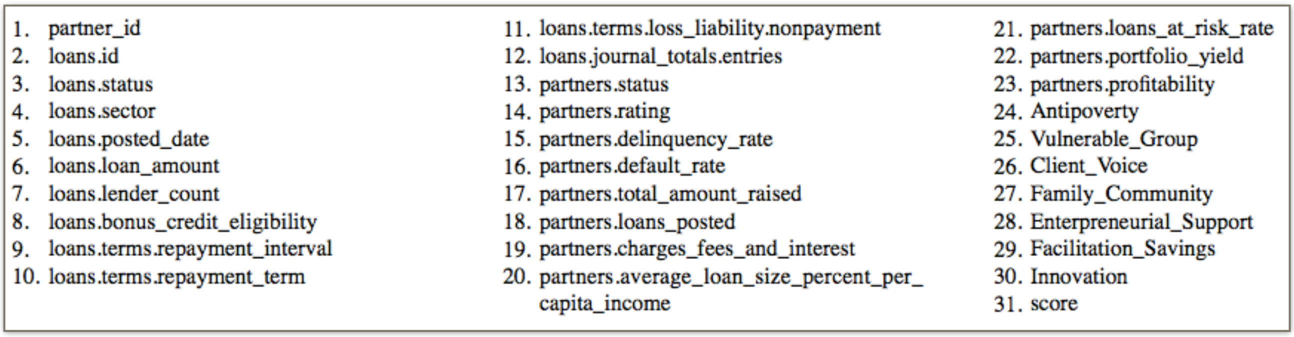
In some cases, it is possible to fill in the missing observations with means (for numeric entries) or modes (for categorical entries). Other variables, such as the text descriptions, cannot be easily inferred. Hence, those observations are omitted from the final dataset before the regression analyses take place. Once all missing observations have been corrected, all the variables are double checked to ensure correctness of the measurement levels. Measurement levels are also changed where appropriate.

4.6. Subsetting the Dependent Variable

In order to use the status of a loan as a Dependent Variable (see Chapter 5), it is necessary to select the relevant levels of the factor. Thus, the subset containing only observations with “loans.loan_status” equal to “paid” or “defaulted” is selected. At the end of this process, the dataset contains 14877 observations of 31 variables.

4.7. Final Dataset and Summary Statistics

While **Figure 5** presents the final set of variables in use, some summary statistics are shown in **Table 5**.



1. partner_id
2. loans.id
3. loans.status
4. loans.sector
5. loans.posted_date
6. loans.loan_amount
7. loans.lender_count
8. loans.bonus_credit_eligibility
9. loans.terms.repayment_interval
10. loans.terms.repayment_term
11. loans.terms.loss_liability.nonpayment
12. loans.journal_totals.entries
13. partners.status
14. partners.rating
15. partners.delinquency_rate
16. partners.default_rate
17. partners.total_amount_raised
18. partners.loans_posted
19. partners.charges_fees_and_interest
20. partners.average_loan_size_percent_per_capita_income
21. partners.loans_at_risk_rate
22. partners.portfolio_yield
23. partners.profitability
24. Antipoverty
25. Vulnerable_Group
26. Client_Voice
27. Family_Community
28. Enterpreneurial_Support
29. Facilitation_Savings
30. Innovation
31. score

FIGURE 5: NAMES OF THE FINAL SET OF VARIABLES USED IN THE ANALYSES

Clean Dataset n=41488	Mean	SD	Min	Max
Loan Amount	748.8	711,91	25.0	10000.0
Repayment Term (months)	11.67	4.78	2	52
Number of Lenders	22.89	21.89	1	331
Partners' Default Rate (%)	1.72	4.78	0	51.96

TABLE 5: DESCRIPTIVE STATISTICS

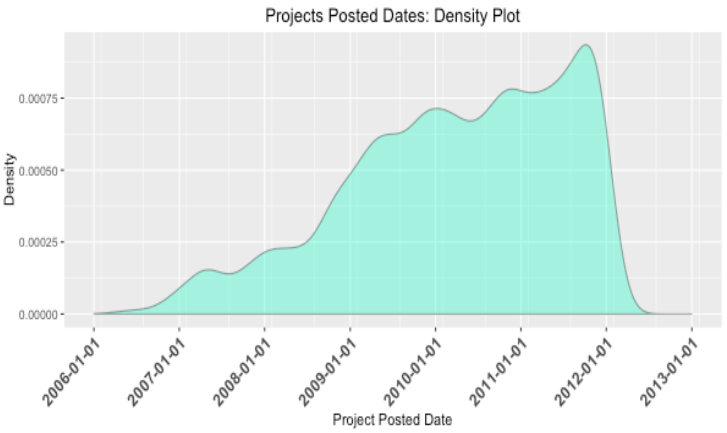


FIGURE 6: DENSITY OF OBSERVATIONS BY POSTED DATE

Table 5 shows that the average loan requested by Kiva borrowers in the dataset is $748.8, higher than the average loan size listed on Kiva website to date ($408.81, 1/06/2016). The average number of lenders per project is 22.89, with a maximum lender pool of 52 users. To date, 1449084 Kiva users have funded a loan on the website. Regarding Field Partners, the sample shows an average default rate of 1.72% over 174 unique Partners. This figure is near to the up-to-date average published on the Kiva website: 1.3% over 299 Field Partners. Another interesting figure (**Figure 6**) is the distribution of loans according to the date they were posted: in the dataset of this study, the majority of the loans are concentrated between the second half of 2008 and the first half of 2013, despite having included also projects from December 31st, 2013 to January 1st, 2016. This sample is still useful for the purposes of this research, because there is a sufficient project density both before and after the introduction of Social Performance Badges (December 11th, 2011).

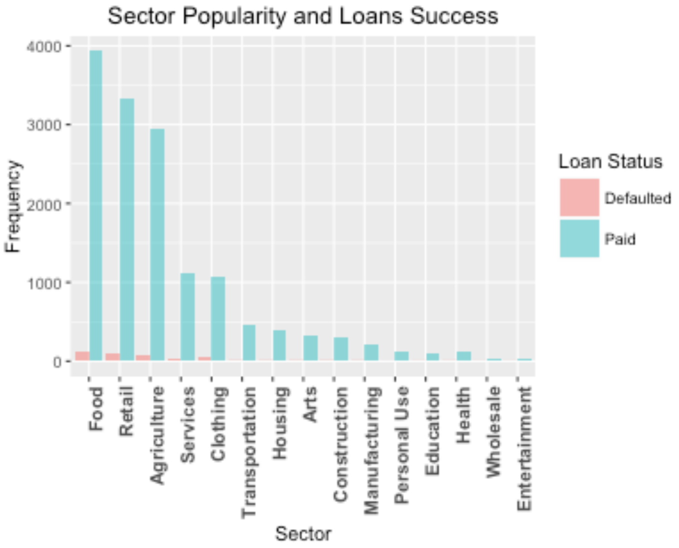


FIGURE 7: DENSITY OF OBSERVATIONS BY PROJECT SECTOR

For what concerns the use of the loans, the most popular sector for funding projects is the Food sector, immediately followed by Retail and Agriculture. This may be due to the fact that the projects are started in developing countries, by people seeking for a long-term source of sustainment (**Figure 7**). In **Figure 8**, it also appears that the amount requested by the borrower doesn't show a clear relationship with loan success. Since the loan amount distribution is skewed towards the left, the loan amount on log scale is used for illustrative purposes. The highest density of projects corresponds to an amount of 1000 (shown on log scale), which is also the point at which the number of defaulted projects is highest. The rating of the Field Partners (**Figure 9**), instead, seems to play a role in discriminating between successful and unsuccesful projects. The majority of defaulted loans appears to be concentrated in the area corresponding to a partner's rating between 0 and 1.

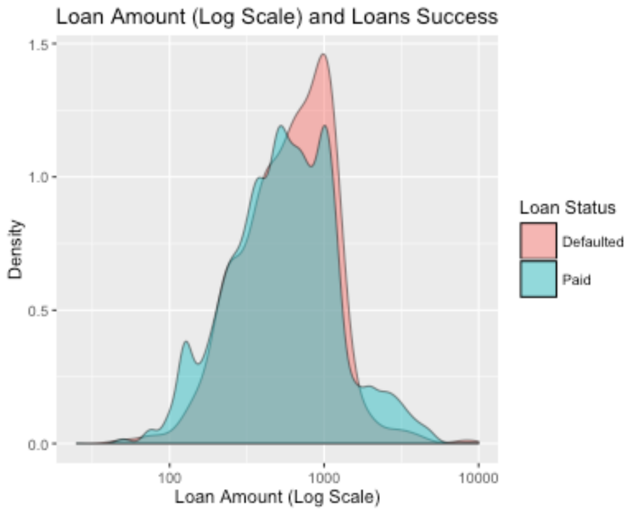


FIGURE 8: DENSITY OF OBSERVATIONS BY LOAN AMOUNT (LOG SCALE)

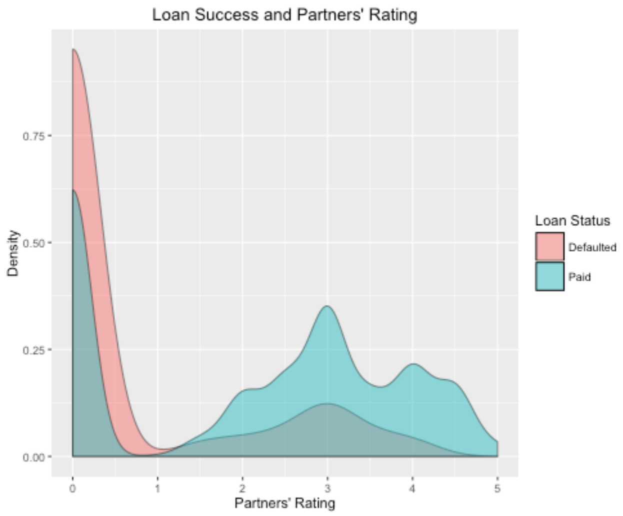
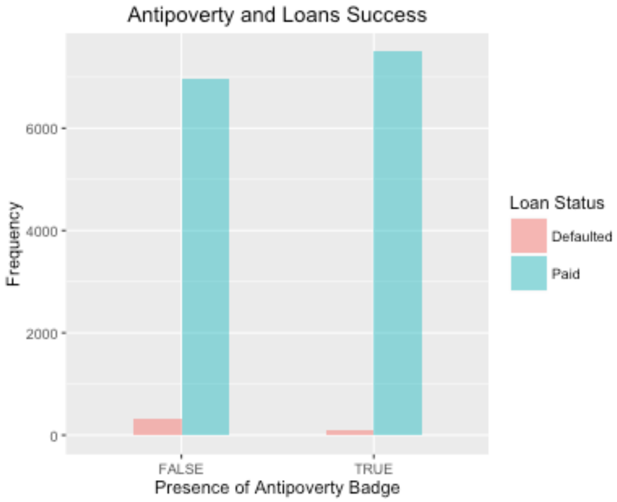


FIGURE 9: DENSITY OF OBSERVATIONS BY PARTNERS' RATING

4.8. Social Performance Badges Statistics

Before assessing the effects of the presence of one or more Badges on the success of a loan, data can be plotted for an initial overview of interesting relations. **Figure 10** shows that the presence of the Antipoverty Focus badge seems related to more successful loans. The same result is obtained also when plotting the remaining badges, as shown in Appendix.



Antipoverty Badge	Defaulted	Repaid
FALSE	317	6966
TRUE	97	7597

FIGURE 10: BARPLOT OF OBSERVATIONS BY PRESENCE OF ANTIPOVERTY BADGE AND CROSS-TABLE OF FREQUENCIES

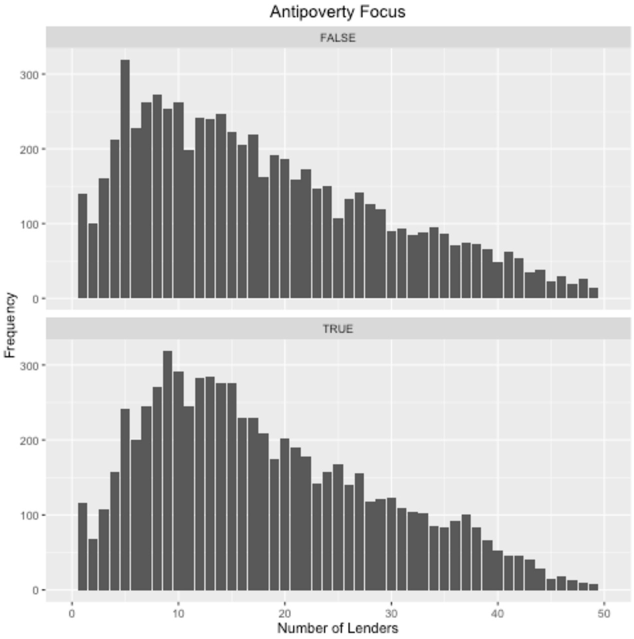


FIGURE 11: NUMBER OF LENDERS AND ANTIPOVERTY FOCUS

For what concerns the number of lenders, the presence of a badge doesn't seem to affect considerably the participation. As **Figure 11** shows, when the Antipoverty Focus is present, the distribution of lenders shifts slightly towards the right (centered at 9 instead of 5), but no major changes appear based on the presence of the symbol. The same holds also for the other Performance Badges (Appendix).

As a final remark, from this preliminary analysis, data appear as aggregated by loan and grouped by loan id. This means that it is only possible to know the loans that have been made and chosen in aggregate terms. So, while it is possible to analyze loan success across different loans, assessment within individuals is not possible. Nonetheless, these data could be useful because, while providing some unique point of reference (i.e. loans.id, loans.partner_id), they also give insights about some determinants of success of the loan and the magnitude of the project at stake (loans.loan_amount).

4.9. Text Descriptions

The text descriptions used for sentiment inspection can be found under the column "loans.description.texts". For the purposes of the present research, only text descriptions in English are taken into consideration. In fact, the sentiment analysis assumes the use of an external database of words, each with positive or negative valence. This research uses external sets of words in English. As a consequence, descriptions in other languages are excluded, as already mentioned in the data cleaning section. Two examples of text descriptions from the analyzed dataset are shown in **Figure 12.**

dataset $loans.description.texts.en[[1]]	The members of the group Sope Khadim, being part of the eco-village Thiaroye, intend to fight against idleness and to take charge of their own destiny. This is why the women got organized into a group. They prefer to be active in the socio-economic advancement of all its members by participating in the development of their community. The objective of this loan is to bolster their business of processing local grain. This is work they've already mastered, and it is centered on purchasing local grain (millet, corn, sorghum) and then proceeding to process it. The sale of the resulting products \"thiacri\" and \"araw\" (the Senegalese names) will allow the loan to be repaid without much difficulty. The turnover of these products pose no problem because the quantity to be sold is less than the level of demand. This loan will allow them to improve the quality and quantity of the goods to be sold.
dataset $loans.description.texts.en[[1990]]	Rafael lives in San Miguel and he has an upholstering and embroidery business. He lives with his 6-year old daughter and his wife, both of whom are financially dependent on him. He starts work very early in the morning and he has two employees. Rafael's business has the advantage that where he has his shop there are no other similar businesses around. He needs to purchase materials, which is why he requests a loan. He needs fabrics, threads, plastics, glue, and other materials to strengthen his business' inventory and to provide his clients with a better level of service. He dreams of being the owner of various businesses and to grow as a small business owner.

FIGURE 12: EXAMPLES OF TEXT DESCRIPTIONS (OBSERVATIONS 1 AND 1990)

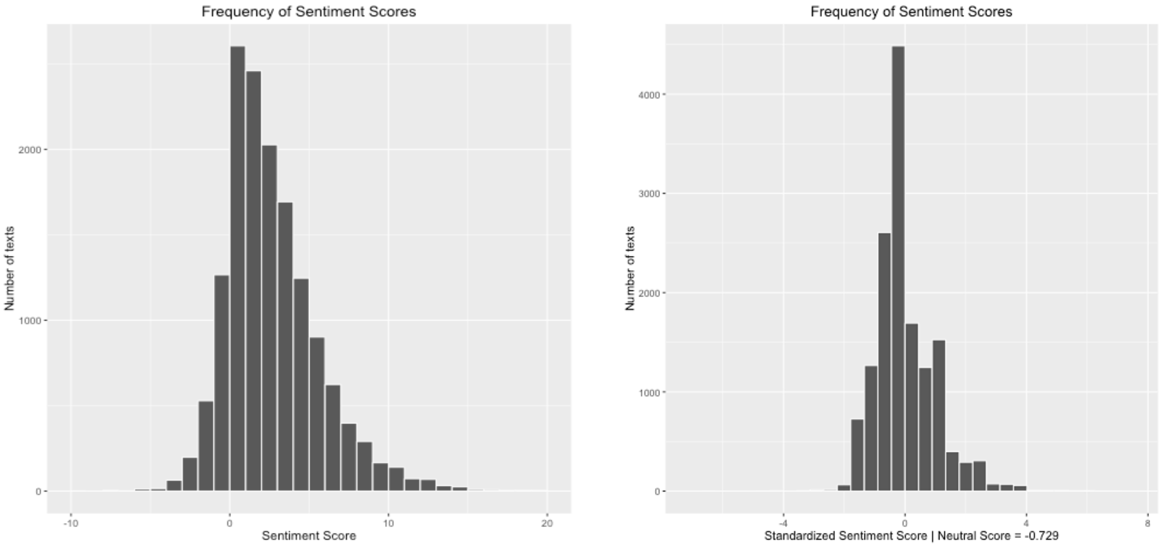
The text descriptions contain several words that are, by themselves, emotionally loaded. Examples from the first observations are “development”, “improve” (positive valence) and “fight”, “idleness” (negative valence). With the help of text mining and sentiment analysis, a systematic detection of emotionally loaded words can be performed. The output is a sentiment score, which is the difference between positive and negative words. Sentiment scores below zero are associated with texts considered inherently negative. Inherently positive texts have an associated sentiment score above zero. Neutral texts have a null sentiment score (Breen, 2012).

4.10. Text Preparation

In order to complete the dataset with the emotional scores of the text descriptions, the text must undergo a cleaning phase. The descriptions must then be cleaned of noisy elements, such as punctuation, special characters, numbers and unnecessary spaces. According to the type of analysis and package used, this phase can occur either before running the sentiment analysis or can be included in the scoring function.

4.11. Assigning Sentiment Scores

The first thing to do after cleaning the text is to import a database of positive and negative words in English. For this research, a list of positive and negative opinion words compiled by Hu and Liu (2004) is used. The list is accessible at the link http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar and is divided into two text files, respectively of positive and negative words. Successively, scores are assigned to the words according to their polarity. A few words specific to the Kiva case are added to the wordlist. Examples of such words are "poverty", "refugees", "abusive", "flawed", "slave", "exploit", "sick", "tired", "destroy" (negative valence) and ‘improve’, ‘soon’, 'hope', 'wish', 'management', 'morality', 'employed' (positive valence). Afterwards, it is necessary to apply the "score.sentiment" function (Breen, 2012), which results in a dataframe containing the analyzed text. The function assigns a score to indicate the valence of the text description. The "scores" dataset is then merged with the relative subset of the original dataframe in order to understand how this scores relate to other variables and proceed with further analyses. Finally, the different subsets, merged with the relative scores, are stacked back to recreate the final dataset. From an initial overview of the data, it appears that text descriptions are positively skewed and the majority of texts are neutral (**Figure 13A**). For visualization and descriptive purposes, the variable is thus standardized, such that observations have mean equal to 0 and standard deviation equal to 1 (**Figure 13B**).



FIGURES 13A AND 13B: FREQUENCY OF SCORES AND OF STANDARDIZED SCORES

The standardized view of the sentiment scores gives a clearer picture of the scores distribution. Given that a neutral score corresponds to a standardized value of -0,73, the distribution is still slightly positively skewed, but the skew is much less evident. Neutral texts are still the majority, but it is more evident how also negative scores are numerous in the dataset. **Figure 14** shows how the sentiment scores are distributed when grouped by the status of the loan. The majority of the defaulted loans (red shaded area) are associated with a neutral text description. Repaid loans are spread between positive and negative texts, but overall concentrated towards the positive side of the density plot.

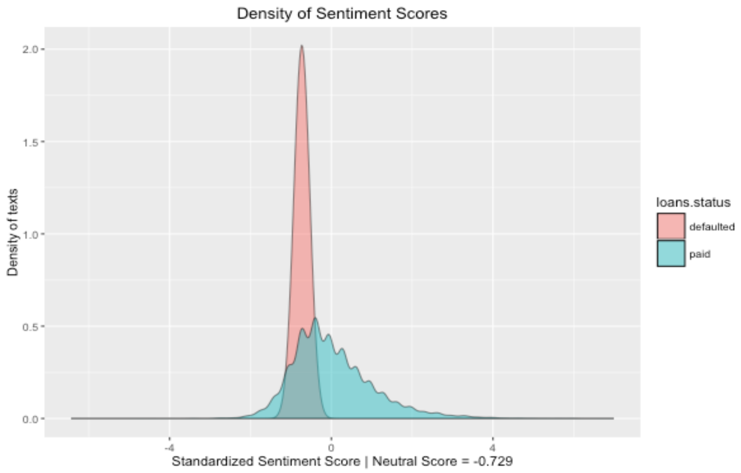


FIGURE 14: DENSITY OF STANDARDIZED SENTIMENT SCORES, GROUPED BY LOAN STATUS

For what concerns the Social Performance Badges, in most cases sentiment does not vary remarkably according to their presence. In the case the badge is more absent than present, the distribution of the sentiment scores remains similar (**Figure 15**).

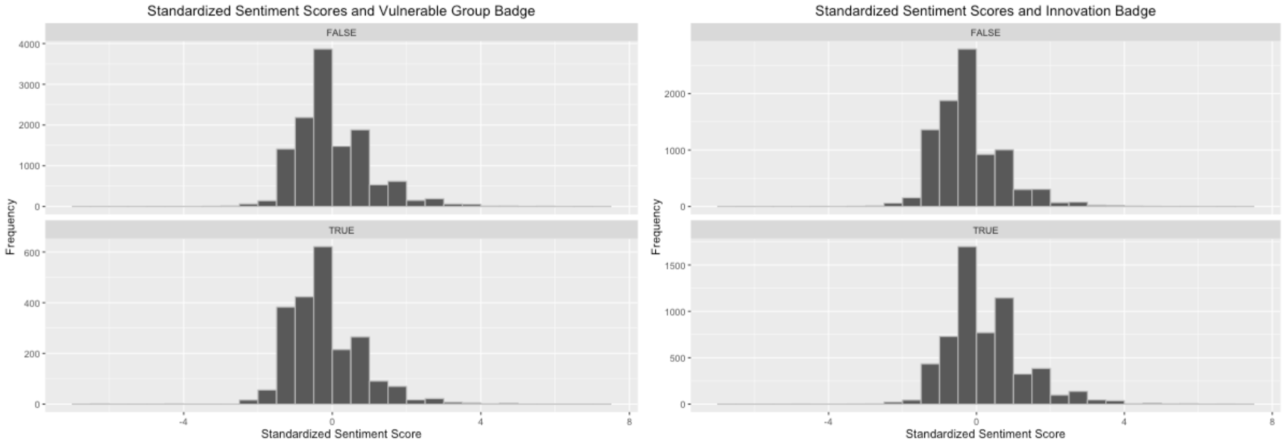


FIGURE 15: DISTRIBUTION OF SENTIMENT SCORES BASED ON PRESENCE OF SELECTED SOCIAL PERFORMANCE BADGES

Only in some cases, the distribution may change in favour of more positive or negative scores when the badge is either present. Figure 15 shows the example of the Innovation Badge. It can be notice that, in presence of the Innovation badge, the distribution is subject to a change in the skew from negative scores to positive scores. Differently, in presence of the Vulnerable Group badge, the distribution of scores seems to keep the same shape. When also the success of the project is added to the picture, it appears that the relationship between sentiment and score reflects the one pictured in Figure 14 above, regardless of the presence of a Social Performance Badge (**Figure 16**).

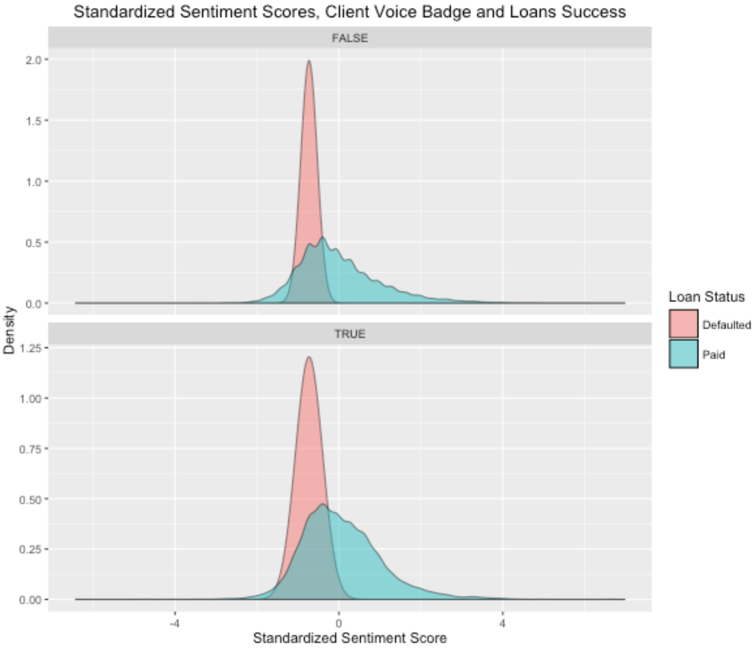


FIGURE 16: DENSITY OF SENTIMENT SCORES BY LOAN STATUS AND CLIENT VOICE BADGE

4.12. Note About the Social Performance Badge Variables

The presence of one or more Social Performance Badges in a loan project is indicated in the present dataset by 7 logical variables, one for each badge, which take the value TRUE when the badge is present in the given project. It is important to notice, however, that the badges are not project-specific: they are assigned every year, following a due diligence process, to the Field Partners which showed a committment to certain social performance areas (see Section 1 of the present research). This means that, even though a project has a certain badge, it is neither necessarily true that that badge will be maintained in the future, nor that it was already present in the past. Thus, for the purposes of this research, the assumption is made that the partners showed committments to the same social performance areas over time. This assumption is strong and is expected to be confirmed or disconfirmed in future research, with data collected over different periods of time.

5. First Regression Analysis

5.1. Dependent Variable: Possibilities and Choice

In the original research design, there were two options for the dependent variable: either the successful repayment of the loan or the loans funded amount. The original idea of using the loan funded amount would have put the research in line with existing literature about decision-making in microfinance, charitable giving and prosocial lending, whereas loan success would have been in line with research about predictive models of loan repayment. The options and the implications of the choice are described as follows:

- *Option 1: Successful Repayment of the Loan*

  Choosing the successful repayment of the loans as dependent variable for the research analyses would have two major implications: first, it would put the research in line with existing literature about prediction of project success in microfinance and prosocial lending. Second, it would mean to analyze data using a binomial logistic regression, since the variable would be categorical, with two levels. The successful repayment would hence be coded as a variable with two levels ("loans.status = paid" and "loans.status = defaulted"). All the intermediate steps of the funding project would be excluded from the analysis. The "paid" status, in fact, implies the occurrence of all the previous steps in the funding request (i.e. the loan has undergone fundraising, was funded and repaid). Moreover, "loans.delinquent" variable would be excluded from the analysis, although it may be an indicator of loan success. Broadly speaking, delinquency may be thought of as an indicator of loan success. Loans, in fact, become delinquent if either the borrower or the Field Partner falls behind on making repayments. Which means that, while the parties responsible for handling repayments are failing to comply with their duties, in principle the project was fully funded. However, since the loan becomes delinquent, it is unclear whether it will eventually develop into default or the delinquency is only a way to take time before the repayment. In conclusion, this variable would be excluded from the analysis, given that the benefits of its inclusion would not offset the increase in complexity that would result.

- *Option 2: Funded Amount*

  For what concerns the funded amount, two more implications would arise from the choice of this variable as the dependent variable. The first implication would be to include the present research in the field of studies about decision-making process of the lender and not about the success of the projects. The second consequence would be to analyze the data with linear regressions, since the dependent variable would be numerical. In the use of the funded amount as a dependent variable, however, the posting date of the loans represents an obstacle. The loans in the dataset are, in fact, posted at different points in time. This means that loans that have been posted earlier are also likely to have systematically higher funded amount. In order to use the funded amount as a dependent variable, a careful selection must be made in order to have subsets of data where all the loans are posted on the same - or reasonably near - dates, in order to avoid systematic errors. The selection of relevant posted dates would be based on two main criteria. First, the dates would be the most popular within the dataset (selected on the basis of the mode), in order to gather as many observations as possible. Second, the dates would be picked from both before the introduction of Social Performance Badges (11-12-2011) and after, to allow for comparisons.

After trying to build a dataset fulfilling the criteria in Option 2, it became clear that subsets of loans with same posted date contained too few data for the results to be considered externally valid. Even after appending more snapshots to the final dataset (increasing from 20000 to 40000 observations), the subsets had still an insufficient number of observations. Another problem was that amounts lended to an expired loan are automatically refunded to the lender ("all-or-nothing policy"). This makes it even more challenging to consider the funded amount as a sound outcome variable. Following this findings, the analyses are eventually performed on the final dataset, using "loans.status" as a dependent variable. This choice implies the two consequences highlighted in Option 1: positioning this research in line with literature about prosocial lending projects success, and the use a binomial logistic regression to classify loan outcomes.

5.2. Predictors and Model 0

Among the 31 variables that make up the final dataset, 27 are used as predictors in the logistic regression model that will follow. The independent variables in the dataset include both loan-specific and partner-specific measures, in order to assess the impact of these different categories on the loan success. Among the predictors, there are also the presence of Social Performance Badges and the Sentiment Score of text descriptions, previously added to the dataset (Section 4). The full list of 27 predictors is shown in **Table 6**.

Loan-specific Variables	**Partner-specific Variables**	**Social Performance Badges**	**Sentiment Analysis**
loans.sector (all levels)	partners.status (all levels)	Antipoverty	score
loans.loan_amount	partners.rating	Vulnerable_Group	
loans.lender_count	partners.delinquency_rate	Client_Voice	
loans.bonus_credit_eligibility	partners.default_rate	Enterpreneurial_Support	
loans.terms.repayment_interval (all levels)	partners.total_amount_raised	Family_Community	
loans.terms.repayment_term	partners.loans_posted	Facilitation_Savings	
loans.terms.loss_liability.nonpayment	partners.charges_fees_and_interest	Innovation	
loans.journal_totals.entries	partners.average_loan_size_percent_per_capita_income		
	partners.loans_at_risk_rate		
	partners.portfolio_yield		
	partners.profitability		

TABLE 6: FINAL LIST OF PREDICTORS AFTER SELECTION OF DEPENDENT VARIABLE

Notice that some of the predictors in Table 6 are stored as factors. Assuming that a factor has *n* levels, in the regression output, each of the *n - 1* levels will be represented as a predictor, while the *n*th will be used as a baseline. The presence of factors makes the total number of predictors (including factor levels) grow from 27 to 43.

The model is approximated to the following logistic function:

$$\hat{p}(X) = \frac{\exp(\beta_0 + \beta'X)}{1 + \exp(\beta_0 + \beta'X)} \tag{1}$$

Where $\beta'X$ is the linear combination of the 27 predictors and their betas.

The expression $\beta_0 + \beta'X$ in formula (1) corresponds to the Log Odds of success, whose formula is:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta'X \tag{2}$$

Formula (2) allows to model the predictors and the outcome variable as a linear function of X. By applying the exponential function to formula (2), it is possible to obtain the Odds of success, whose formula is:

$$\frac{p(X)}{1 - p(X)} = \exp(\beta_0 + \beta'X) \tag{3}$$

Formula (3), representing the odds that a loan will be repaid, allows for the interpretation of the coefficient estimates. According to it, a 1-unit increase in X corresponds to an increase in the odds of loan repayment by exp(β) units. Odds can take any value between 0 and ∞. Values of the odds close to 0 and ∞ translate respectively in very low and very high probabilities of loan repayment (Gareth, 2013). In the present research, the average effect of a 1-unit change in a predictor on the odds of success, calculated as exp(β), is reported along with the results of the logistic regressions.

5.3. Model 0: Results

The estimated coefficients, standard errors and average effect on odds from the logistic regression process are presented in **Table 7**.

Coefficients	**Estimate**	**Std. Error**	**z value**	**A.E.O.**
Intercept	16,877	7E+03	0,003	-
loans.sector				
Agriculture (baseline)	0,000	0,000	0,000	0,000
Arts	-0,186	0,609	-0,306	0,830
Clothing	-1,224	0,284	-4,308	0,294
Construction	-0,053	0,564	-0,095	0,948
Education	0,022	0,710	0,030	1,022
Entertainment	-2,149	0,845	-2,543	0,117
Food	-0,673	0,230	-2,919	0,510
Health	-0,650	0,616	-1,056	0,522
Housing	0,343	0,508	0,675	1,409
Manufacturing	-0,348	0,716	-0,486	0,706
Personal Use	0,038	1,087	0,035	1,039
Retail	-0,906	0,242	-3,744	0,404
Services	-0,556	0,313	-1,774	0,574
Transportation	-0,207	0,477	-0,435	0,813

Coefficients	Estimate	Std. Error	z value	A.E.O.
Wholesale	-1,395	0,875	-1,595	0,248
loans.loan_amount	0,001	0,000	3,133	1,001
loans.lender_count	-0,043	0,014	-3,177	0,958
loans.bonus_credit_eligibility				
True	-0,873	0,306	-2,851	0,418
loans.terms.repayment_interval				
End-of-the-Term (baseline)	0,000	0,000	0,000	0,000
Irregularly	1,292	0,528	2,447	3,640
Monthly	1,165	0,410	2,840	3,206
loans.terms.repayment_term	-0,134	0,017	-8,117	0,875
loans.terms.loss_liability.nonpayment(partner)	1,823	0,185	9,846	6,192
loans.journal_totals.entries	-0,540	0,070	-7,694	0,583
partners.status				
active (baseline)	0,000	0,000	0,000	0,000
closed	1,261	0,656	1,922	3,527
inactive	13,272	4E+02	0,036	6E+05
paused	-1,331	0,662	-2,011	0,264
partners.rating	0,981	0,234	4,191	2,667
partners.delinquency_rate	-0,027	0,011	-2,346	0,974
partners.default_rate	-0,204	0,011	-18,303	0,816
partners.total_amount_raised	0,000	>0,001	0,110	1,000
partners.loans_posted	0,000	>0,001	0,167	1,000
partners.charges_fees_and_interest				
True	-12,718	7E+03	-0,002	0,000
partners.average_loan_size_percent_per_capita_income	-0,012	0,002	-7,831	0,988
partners.loans_at_risk_rate	0,033	0,010	3,247	1,033
partners.portfolio_yield	-0,011	0,007	-1,612	0,989
partners.profitability	0,017	0,018	0,996	1,018
Antipoverty	-0,447	0,239	-1,873	0,640
Vulnerable_Group	-0,656	0,236	-2,773	0,519
Client_Voice	0,169	0,257	0,658	1,185
Family_Community	1,214	0,275	4,421	3,366
Enterpreneurial_Support	0,999	0,286	3,493	2,716
Facilitation_Savings	0,139	0,190	0,730	1,149
Innovation	-0,649	0,258	-2,517	0,522
score	0,457	0,035	13,053	1,580

A.E.O.: Average Effect on the Odds of success = e^β

Chi-square: 1850.619

Df: 43

Associated p-value: 0.000

TABLE 7: RESULTS FROM LOGISTIC REGRESSION (MODEL 0)

5.4. Discussion of the Regression Results

5.4.1. Expectations Confirmed: Loan Specific Variables

Some of the positive and negative coefficient estimates reflect the research expectations. For what concerns the loan-specific predictors, when the liability of non-payment is on the partner (instead of the lender), the project has higher chances of being repaid. In this case, the odds of loan success multiplicatively increase by more than 6 units. The positive coefficient estimate for the loan amount (β_{LA} = 0,001) disconfirms the findings from both Kumar (2007) and Everett (2015), with a 1-unit increase in loan amount associated with a 1-unit multiplicative increase in the odds of loan repayment. Also, differently from what stated by Field and Pande (2008), repayment terms have a significant effect on chances of loan repayment. In fact, a 1-unit increase in the repayment term multiplicatively decreases the odds of success - in particular, a project with 1-unit higher repayment term has 0,87 odds of being repayed.

5.4.2. Expectations Confirmed: Partner Specific Variables

For what concerns partner-specific predictors, higher ratings of the field partners increase the likelihood of the loan being repaid. Specifically, a 1 unit increase in partners' rating multiplicatively increases the log odds of loan success by more than 2,5 units. Field Partners on "paused" status have a more negative impact on loans success than active partners. Furthermore, if a Partner has relatively higher delinquency rate increases the chance of loan default: a 1 unit increase in partners' delinquency rate multiplicatively decreases the odds of loan success down to 0,97. Also, if the average loan size as percent of per capita income disbursed by the Field Partner increases by 1 unit, the odds of loan success multiplicatively decrease to 0,98. Probably, higher amounts at stake are exposed to additional sources of risk. Reasonably, if partners' default rate increases by 1 units, the odds of loan success multiplicatively decrease to 0,82. The positive effect of the increase in profitability rates of field partners is in line with expectations and common sense. The remaining two coefficient estimates are both partner-specific, and are the "charges fees and interest" and the "portfolio yield" variable. The portfolio yield is defined as the average interest rate and fees that Kiva borrowers pay to the Kiva Field Partner. Thus, it is reasonable that, as the interest rate and fees associated to a loan increase, the loan is more likely to default. This may be due to the fact that the loan is more risky than others, and hence requires the implementation of increased interest rates. The same line of reasoning can be applied to the "charges fees and interest" variable.

5.4.3. Expectations Confirmed: Social Performance Badges and Sentiment Scores

Regarding the effect of sentiment scores, a project with a more positively valenced text description has higher chances of being repaid: a 1-unit increase in the sentiment score multiplicatively increases the odds of success by 1,57. This confirms the results from Gao and Lin (2015), and validate the research expectation that these findings could be extended to the environment of an online platform other than Kiva.org. Another result that validates research expectations is that the presence of the Family and Community and Enterpreneurial Support social performance badges makes a loan more likely to be repaid, as opposed to their absence. The multiplicative increase in the odds of success is by 3,37 and 2,72 units respectively. The presence of the Client Voice and Facilitation of Savings social performance badges is also associated with an increase in the chances of loan repayment (β_{CV} = 0,169 and β_{FS} = 0,139).

5.4.4. Expectations Disconfirmed: Loan Specific Variables

While all of the above makes sense according to both prior research and expectations, other coefficient estimates appear counter-intuitive. Irregular repayment terms increase the odds of loan success more than Monthly or End-of-the-Term terms, which appears to be against common risk-aversion. The coefficients associated to loan sectors appear to be mostly negative. In fact, projects in the Retail, Clothing, Entertainment, Food and Service sectors decrease the odds of loan success, when compared to Agriculture projects. Furthermore, the more lenders to a project, the less likely it is to be repaid: a 1 unit increase in the number of lenders slightly decreases (β_{NL}= -0,043) the log odds of loan success, with odds equal to 0,96. An increase by 1 unit of loan at risk rate is associated with a positive coefficient estimate (β_{LRR}= 0,033). It would be more reasonable to expect a negative estimated beta coefficient. Also the bonus credit eligibility is associated with a negative coefficient estimate (β_{BCE}= -0,873). The bonus credit is a $25 bonus earned for each new user invited on Kiva by the lender. This eligibility subject to condition may be the reason of the negative coefficient estimate. Also a 1-unit increase in the number of project journal entries multiplicatively decreases the odds of success: repayment is only 0,58 as likely as default. This is against the expectation that more detailed and updated projects are expected to succeed. There are no details about the content of the journal entries.

5.4.5. Expectations Disconfirmed: Partner Specific Variables

For what concerns partner-specific variables, if a partner has a "closed" status instead of "active", the chances of the loan being repaid increase. However, there may be some time-related biases in this coefficient estimates that the model failed to capture. For example, successful past loans that were administered by a partner with status "closed" at the moment of the latest data snapshot update. Partners with "inactive" status have a strongly positive coefficient estimate, which is against expectations. However, the same considerations apply as for the "closed" partner status. The effects of the total amount raised and the number of loans posted are instead very small, and, in any event, their impact would have been limited.

5.4.6. Expectations Disconfirmed: Social Performance Badges

Some results concerning the Social Performance Badges also disconfirm research expectations. The presence of either the Vulnerable Group, the Innovation or the Antipoverty social performance badges appears to have a negative impact on loan repayment when compared to their absence. In fact, the presence of the badges multiplicatively decreases the odds of loan success, with values between 0,52 and 0,64.

5.5. Prediction and Model Accuracy

To asses the predictive accuracy, the model is trained using 10-fold cross validation and then predicted on the test data. **Table 8** shows the predictive power of the model.

	Glm Predictions*	
Observed Outcome	**Defaulted**	**Paid**
Defaulted	141	53
Paid	273	14410

*p_0 = 50% , n = 14877

TABLE 8: CONFUSION MATRIX WITH ARBITRARY THRESHOLD - MODEL 0

Assuming an arbitrary prediction threshold (p_0) of 50% for loan success, the model correctly predicts the repayment of a loan 97.81% of the times.

It is possible to use the test set also to look for an optimal threshold (p_0). Let p_0 vary along a range between 0.001 and 0.999, with length 100; the optimal p_0 is obtained when the minimum overall error rate is reached. **Table 9** shows some statistics of the error rates and the prediction thresholds at which the minimum overall error rate is achieved.

	Overall Error Rate	**False Positive**	**False Negative**	**True Positive**	**Threshold (p_0)**
Average	0,0362	0,5907	0,0204	0,9796	50,0%
Minima	0,0210	0,4903	0,0075	0,9925	67,6%
	0,0210	0,4807	0,0078	0,9922	68,6%

TABLE 9: OPTIMAL PREDICTION THRESHOLDS AND SUMMARY STATISTICS - MODEL 0

There are two threshold at which the minimum overall error rate is reached. The choice depends on which probability is the model predicting and the field of application of the model. In this case, the model predicts the probability of a loan being successfully repaid. It is important to remember that field of application is prosocial lending — where lenders might be socially involved but financially risk-averse, and might prefer loans with higher percentages of predicted success. Hence, it is safe to assume that the threshold for p_0 associated with the lowest proportion of false positives is to be preferred. The choice of p_0 is crucial, because only those loans which had a predicted probability of repayment higher than p_0 will be classified as repaid, while all the rest is interpreted as defaulted. Based on this outcome, the prediction is repeated, this time assuming $p_0 = 0.686$ (**Table 10**).

	Glm Predictions*	
Observed Outcome	**Defaulted**	**Paid**
Defaulted	215	113
Paid	199	14350

*p_0 = 68,6% , n = 14877

TABLE 10: CONFUSION MATRIX WITH OPTIMAL THRESHOLD - MODEL 0

By changing the optimal threshold p_0 by 0.186, the model correctly predicts loan repayment 97.90% of the times - as opposed to 97.81% with (p_0) = 0.5.

6. Variable Selection and Regularization

The first regression analysis was performed on the full set of 27 independent variables, resulting in 43 predictors. It may be useful to look for a model which has similar or better predictive power, but uses less variables. To reach this objective, a lasso regression for variable selection purposes is applied. The lasso regression applies a shrinking penalty on the size of the parameters, such that certain parameters are automatically set to 0 when the tuning parameter (λ) is sufficiently large. This allows to perform a parameter selection and to achieve a two-fold objective. First, control for the complexity of the model. Second, have the freedom to choose models in a range from very flexible (lasso λ small) to very inflexible (lasso λ large). After this procedure, the accuracy of the selected model with a reduced number of variables is assessed through K-fold cross validation.

6.1. Creation of Suitable Dataset

In order to run the lasso regression, categorical variables are first transformed into factors. Then, a dummy variable matrix of predictors is created. The matrix is filled with the categorical entries and with the continuous predictors and then is passed to the general linear model. The list of categorical and continuous predictors can be found in **Table 11**.

Categorical Variables		**Continuous Variables**	
Loan- and Partner-specific	**Social Performance Badges**	**Loan- and Partner-specific**	**Score Sentiment**
loans.sector (all levels)	Antipoverty	loans.loan_amount	score
loans.bonus_credit_eligibility	Vulnerable_Group	loans.lender_count	
loans.terms.repayment_interval (all levels)	Client_Voice	loans.terms.repayment_term	
loans.terms.loss_liability.nonpayment	Enterpreneurial_Support	loans.journal_totals.entries	
partners.status (all levels)	Family_Community	partners.rating	
partners.charges_fees_and_interest	Facilitation_Savings	partners.delinquency_rate	
	Innovation	partners.default_rate	
		partners.total_amount_raised	
		partners.loans_posted	
		partners.average_loan_size_percent_per_capita_income	
		partners.loans_at_risk_rate	
		partners.portfolio_yield	
		partners.profitability	

TABLE 11: PREDICTORS LIST DIVIDED BY MEASUREMENT LEVEL

Another step before fitting the lasso regression is to create a dummy variable for the dependent variable "loans.status" in order to get the value of the best lambda afterwards. For the purposes of this analysis, the dummy is created such that loans.status = 1 if loans.status = "paid" and loans.status = 0 if loans.status = "defaulted".

6.2. Choice of Tuning Parameter with 10 Fold Cross Validation

After fitting the model on a train set, it can be noticed that, depending on the choice of the tuning parameter, some of the 43 coefficients are shrinked to zero (**Figure 17**).

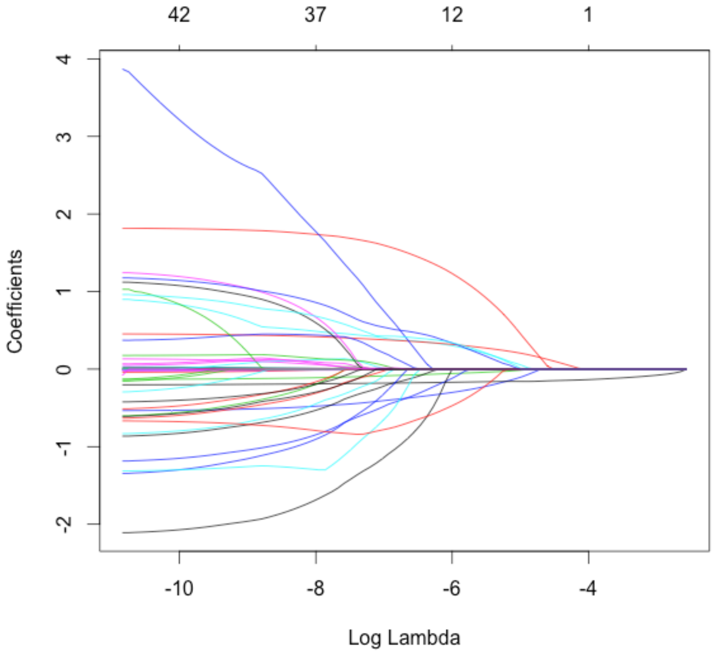


FIGURE 17: PLOT OF COEFFICIENTS SHRINKAGE BASED ON VALUES OF LOG LAMBDA

A 10-fold cross validation is performed to choose the optimal value of the tuning parameter (λ), and then the tuning parameter for which the cross-validation error is smallest is selected. The final values of λ range from 8.033e-02 (df = 0, %Dev = 0.00000) to 3.906e-05 (df = 42, %Dev = 0.25920). The value of the tuning parameter for which the mean cross-validated error is minimized is 0.000174, which corresponds to df = 40 and %Dev = 0.25890.

6.3. Model 1: Fit and Assessment

The logistic model is then fitted again, this time setting the value of the tuning parameter at the optimal 0.000174. **Table 12** shows the results of the procedure, in particular, the predictors and the associated coefficient estimates. Variables penalized during the shrinkage process are assigned an estimate of 0 and highlighted in bold.

Predictor	**Est.Coeff.**
(Intercept)	5,239
loans.loan_amount	0,001
loans.lender_count	-0,024
loans.terms.repayment_term	-0,120

Predictor	**Est.Coeff.**
loans.journal_totals.entries	-0,506
partners.rating	0,537
partners.delinquency_rate	-0,011
partners.default_rate	-0,194
partners.total_amount_raised	0,000
partners.loans_posted	0,000
partners.average_loan_size_percent_per_capita_income	-0,011
partners.loans_at_risk_rate	0,015
partners.portfolio_yield	-0,012
partners.profitability	0,014
score	0,434
loans.sectorArts	**0,000**
loans.sectorClothing	-0,994
loans.sectorConstruction	0,100
loans.sectorEducation	0,120
loans.sectorEntertainment	-1,902
loans.sectorFood	-0,439
loans.sectorHealth	-0,368
loans.sectorHousing	0,452
loans.sectorManufacturing	-0,006
loans.sectorPersonal.Use	0,136
loans.sectorRetail	-0,668
loans.sectorServices	-0,306
loans.sectorTransportation	**0,000**
loans.sectorWholesale	-1,087
loans.bonus_credit_eligibilityTRUE	-0,625
loans.terms.repayment_intervalIrregularly	0,956
loans.terms.repayment_intervalMonthly	0,866
loans.terms.loss_liability.nonpaymentpartner	1,779
partners.statusclosed	**0,000**
partners.statusinactive	2,400
partners.statuspaused	-1,251
partners.charges_fees_and_interestTRUE	**0,000**
AntipovertyTRUE	-0,305
Vulnerable_GroupTRUE	-0,733
Client_VoiceTRUE	0,181
Family_CommunityTRUE	0,990
Enterpreneurial_SupportTRUE	0,768
Facilitation_SavingsTRUE	0,094
InnovationTRUE	-0,402

TABLE 12: RESULTS FROM LASSO REGRESSION (MODEL 1)

As Table 12 shows, only 2 predictors in full (partners.status "closed" and partners.charges_fees_and_interest) and 2 levels of the variable loans.sector are assigned a null coefficient estimate. The final model to use for predictions would thus entail the presence of all the predictors from the initial model (Model 0), with the exception of partners.charges_fees_and_interest, partners.status level "closed" and loans.sector levels "Arts" and "Transportation".

6.4. Predictive Accuracy

Table 13 shows the predictive power of Model 1, based on a 10-fold cross validation and an optimal prediction threshold (p_0) of 68.65%.

	Glm Predictions*	
Observed Outcome	**Defaulted**	**Paid**
Defaulted	209	97
Paid	205	14366

*p_0 = 68,65% , n = 14877

TABLE 13: CONFUSION MATRIX WITH OPTIMAL THRESHOLD - MODEL 1

The confusion matrix shows how, at an optimal prediction threshold (p_0) of 68.65%, Model 1 correctly predicts the repayment of a loan 97.97% of the times. With an improvement in prediction accuracy of 0.07, Model 1 predicts loan repayment slightly more accurately, while using fewer variables than Model 0.

6.5. Comparison Between Model 0 and Model 1

Instead of commenting the lasso coefficients as standalone values, it may be interesting to see how the estimated effects of the coefficients on the probability of loan success have changed, as opposed to the first model (**Table 14**).

Coefficients	**Estimate Model 1**	**Estimate Model 0**
(Intercept)	5,239	16,877
loans.loan_amount	0,001	0,001
loans.lender_count	-0,024	-0,043
loans.bonus_credit_eligibilityTRUE	-0,625	-0,873
loans.terms.repayment_intervalIrregularly	0,956	1,292
loans.terms.repayment_intervalMonthly	0,866	1,165
loans.terms.repayment_term	-0,120	-0,134
loans.terms.loss_liability.nonpaymentpartner	1,779	1,823
loans.journal_totals.entries	-0,506	-0,540
partners.rating	0,537	0,981
partners.delinquency_rate	-0,011	-0,027
partners.default_rate	-0,194	-0,204
partners.total_amount_raised	0,000	0,000
partners.loans_posted	0,000	0,000
partners.average_loan_size_percent_per_capita_income	-0,011	-0,012
partners.loans_at_risk_rate	0,015	0,033
partners.portfolio_yield	-0,012	-0,011
partners.profitability	0,014	0,017
AntipovertyTRUE	-0,305	-0,447
Vulnerable_GroupTRUE	-0,733	-0,656
Client_VoiceTRUE	0,181	0,169
Family_CommunityTRUE	0,990	1,214
Enterpreneurial_SupportTRUE	0,768	0,999
Facilitation_SavingsTRUE	0,094	0,139
InnovationTRUE	-0,402	-0,649
score	0,434	0,457
dummy_Food	-0,439	-0,673
dummy_Services	-0,306	-0,556
dummy_Agriculture	-0,668	0,000
dummy_Retail	-0,668	-0,906
dummy_Construction	0,100	0,631
dummy_Clothing	-0,994	0,406
dummy_Housing	0,452	0,343

Coefficients	Estimate Model 1	Estimate Model 0
dummy_Wholesale	-1,087	-1,395
dummy_Manufacturing	-0,006	-0,348
dummy_Health	-0,368	-0,650
dummy_Personal_Use	0,136	0,038
dummy_Entertainment	-1,902	-2,149
dummy_Education	0,120	0,022
dummy_active	0 (baseline)	0,000
dummy_paused	-1,251	-1,331
dummy_inactive	2,400	13,272

TABLE 14: COMPARISON OF RESULTS FROM REGRESSION ANALYSES: MODEL 0 AND MODEL 1

Except for the one of the Clothing sector, the coefficient estimates have kept the same signs. However, they differ in magnitude: in particular, there seems to be a tendency for all the coefficients towards the zero. Among others, coefficients estimates for some loan sectors (Food, Services, Housing, Wholesale, Manufacturing, Health, Personal Use, Entertainment, Education and Retail) and for the Antipoverty, Client Voice and Innovation badges have increased, meaning that they are likely to have a slightly more positive impact on the odds of loan replayment. Among the decreased coefficient estimates it is possible to notice the intercept of the model, the presence of the Family and Community, Enterpreneurial Support, Vulnerable Group and Facilitation of Savings social performance badges, the sectors Construction and Clothing and the "inactive" status level for the field partners. The sentiment score estimate has also slightly decreased. These predictors are expected to have a slightly less positive (or more negative) impact on the odds of loan repayment compared to Model 0.

6.6. Comparison Between Lasso Regression and Logistic Regression

As an additional attempt to find the best model for the loan classification problem, another logistic regression is performed on the data. This time, the predictors of the new logistic regression would be those selected by the lasso regression performed earlier. The coefficient estimates obtained through this procedure are shown below, in **Table 15**.

Coefficients	Estimate (Linear Regression on Selected Predictors)	Estimate (Lasso Regression)
(Intercept)	5,220	5,239
loans.loan_amount	0,001	0,001
loans.lender_count	-0,043	-0,024
loans.bonus_credit_eligibilityTRUE	-0,873	-0,625
loans.terms.repayment_intervalIrregularly	1,292	0,956
loans.terms.repayment_intervalMonthly	1,165	0,866
loans.terms.repayment_term	-0,134	-0,120
loans.terms.loss_liability.nonpaymentpartner	1,823	1,779
loans.journal_totals.entries	-0,540	-0,506
partners.rating	0,981	0,537
partners.delinquency_rate	-0,027	-0,011
partners.default_rate	-0,204	-0,194
partners.total_amount_raised	0,000	0,000
partners.loans_posted	0,000	0,000
partners.average_loan_size_percent_per_capita_income	-0,012	-0,011
partners.loans_at_risk_rate	0,033	0,015

Coefficients	**Estimate (Linear Regression on Selected Predictors)**	**Estimate (Lasso Regression)**
partners.portfolio_yield	-0,011	-0,012
partners.profitability	0,017	0,014
AntipovertyTRUE	-0,447	-0,305
Vulnerable_GroupTRUE	-0,656	-0,733
Client_VoiceTRUE	0,169	0,181
Family_CommunityTRUE	1,213	0,990
Enterpreneurial_SupportTRUE	0,999	0,768
Facilitation_SavingsTRUE	0,139	0,094
InnovationTRUE	-0,649	-0,402
score	0,457	0,434
dummy_Food	-0,473	-0,439
dummy_Services	-0,356	-0,306
dummy_Agriculture	0,200	-0,668
dummy_Retail	-0,706	-0,668
dummy_Construction	0,146	0,100
dummy_Clothing	-1,024	-0,994
dummy_Housing	0,543	0,452
dummy_Wholesale	-1,195	-1,087
dummy_Manufacturing	-0,148	-0,006
dummy_Health	-0,450	-0,368
dummy_Personal_Use	0,238	0,136
dummy_Entertainment	-1,949	-1,902
dummy_Education	0,221	0,120
dummy_active	-1,260	0 (baseline)
dummy_paused	-2,591	-1,251
dummy_inactive	12,013	2,400

Chi-square of logistic regression: 1850.612
Df: 41
Associated p-value: 0.000

TABLE 15: COMPARISON OF RESULTS BETWEEN MODEL 1 AND ADDITIONAL LOGISTIC REGRESSION WITH FEWER PREDICTORS

As it can be noticed from Table 15, the difference between lasso coefficient estimates and those from logistic regression is very small. In all cases, positive and negative signs are unchanged. The only change recorded is the magnitude of the coefficient estimates, which seem to be smaller in the lasso column. Since performing the additional logistic regression does not entail different results, only the coefficients from the lasso procedure will be taken into account and, from now on, referred to exclusively as the coefficients from Model 1.

7. Social Performance Badges: Before and After the Introduction

Social Performance Badges have been introduced relatively recently to the Kiva interface: more precisely, Kiva announced the launch of the badges on its blog on December 11st, 2011 (JD Bergeron, December 2011). After having developed a model for the overall set of predictors, it may be interesting to assess the impact of certain variables before and after the introduction of the Social Performance Badges. For this purpose, an ad-hoc model is developed on the dataset containing observations older than the date of the introduction of the badges. This dataset has no Social Performance Badge indicators. After that, due to a small sample size, a lasso regression is applied to the remaining part of the data. This set contains observations posted between December 11st, 2011 and January 1st, 2016. The same assumption specified in Section 1, about the unchanged presence of Social Performance Badges over time, holds also in the following sections.

7.1. Model A: Before the Introduction of the Social Performance Badges

In order to assess the effects of all the predictors before the introduction of the Social Performance Badges, the subset of observations with posted dates between January 1st, 2006 and December 11st, 2011 is taken into account. The variables measuring the presence of the badges are dropped (7 variables, from “Antipoverty” to “Innovation”). The dataset built under these contraints has 14281 observations of 26 variables: thus, the majority of the observations falls within this dataset. The full logistic regression from section 5 (Model 0), without the Social Performance Badge predictors, is run on the dataset. The results from the regression analysis are shown in a table in the Appendix and are discussed in the following sections.

7.1.1. Model A Discussion: Differences with Model 0

Table 16 illustrates the positive and negative effects of the predictors on the probability of loan repayment, before the introduction of the Social Performance Badges.

$\beta \geq 0$	$\beta < 0$
loans.sectorHousing	loans.sectorArts
loans.sectorManufacturing	loans.sectorClothing
loans.sectorPersonal Use	loans.sectorConstruction
loans.sectorTransportation	loans.sectorEducation
loans.loan_amount	loans.sectorEntertainment
loans.terms.repayment_intervalIrregularly	loans.sectorFood
loans.terms.repayment_intervalMonthly	loans.sectorHealth
loans.terms.loss_liability.nonpaymentpartner	loans.sectorRetail
partners.statusclosed	loans.sectorServices
partners.statusinactive	loans.sectorWholesale
partners.rating	loans.lender_count
partners.total_amount_raised	loans.bonus_credit_eligibilityTRUE
partners.loans_posted	loans.terms.repayment_term
partners.loans_at_risk_rate	loans.journal_totals.entries
score	partners.statuspaused
	partners.delinquency_rate
	partners.default_rate
	partners.average_loan_size_percent_per_capita_income

β ≥ 0	β < 0
	partners.portfolio_yield
	partners.profitability

TABLE 16: COEFFICIENT ESTIMATES FROM MODEL A DIVIDED BY POSITIVE AND NEGATIVE IMPACT

Instead of considering the coefficients in isolation, the results from Model A are discussed in comparison to the performance of Model 0. When compared to the results of the logistic regression applied to the full dataset, it is possible to notice some changes in the impact of the predictors on the probability of loan repayment. For what concerns the sector of use of the loan, the levels “Education”, “Manufacturing” and “Transportation” have undergone a change in the sign of the coefficient estimate. This means that, when compared to the results which include the Social Performance Badges, these sectors shifted from having a negative impact (“Education”) to a positive impact, or viceversa (“Manufacturing” and “Transportation”). Partners’ profitability also went from having a negative effect on the probability of loan repayment before December 11st, 2011, to a positive impact in general. The relative impact of irregularly and montly repayment terms on the log odds of repayment is still positive, but was less in magnitude before the cutoff date (around 0,8 versus 1,2 in the full dataset). The variable “partners.charges_fees_and_interest” is always true in the dataset under current investigation, hence its effects are not studied in this regression analysis. The effects of the sentiment scores on the chances of loan repayment are roughly the same.

7.1.2. Model A: Accuracy

In order to assess the accuracy of the model, it is subject to a 10-fold cross validation. Given the optimal prediction threshold (p_0) of 63.61%, the model correctly predicts loan repayment 97.97% of the times. The confusion matrix for the model is presented in **Table 17**.

	Glm Predictions*	
Observed Outcome	**Defaulted**	**Paid**
Defaulted	190	82
Paid	209	13800

*p_0 = 63,61% , n = 14281

TABLE 17: CONFUSION MATRIX AT OPTIMAL THRESHOLD- MODEL A

It can be noticed that the number of false negatives (209) is relatively high. This may pose a threat to the usability of the model for prediction purposes. However, given that this model is retroactive and not intended to be used on unobserved data, the threat is not addressed in this research.

7.2. Model B: After the Introduction of the Social Performance Badges

The analysis can be further extended by studying the remaining observations, namely those projects that were posted after the introduction of the Social Performance Badges. The dataset in use, with the most recent projects, is composed by 580 observations of 46 variables. Some of the variables show singularities: the loss liability of non-payment in the dataset is always on the lender and there is no project with sector equal to “Wholesale”. For practical reasons, these variables are set to 0. The small size of the dataset represents an obstacle to the implementation of a logistic regression. When calling for a logistic regression, the model produces a perfect separation of the predictions, almost certainly due to low amount of obervation points and the relatively high number of variables. For this reason, a penalized logistic regression, the lasso, is preferred.

7.2.1. Model B: Discussion of Results

The results from the analysis are presented in the Appendix. The lasso regression penalized many of the 44 variables for this dataset. Among the non-penalized variables, the sentiment score of the text descriptions keeps its positive effect on the chances of loan repayment. Unexpectedly, all but one of the Social Performance Badges were penalized by the lasso regression, producing little interpretable results. The Client Voice is the only badge with a non-zero coefficient - more specifically, it has a negative estimate. This means that, after its introduction, in presence of the badge, the loan has less chances to be repayed.

7.2.2. Model B: Accuracy

Table 18 shows the confusion matrix associated with Model B.

	Glm Predictions*	
Observed Outcome	**Defaulted**	**Paid**
Defaulted	5	0
Paid	10	565

*p_0 = 69,66% , n = 580

TABLE 18: CONFUSION MATRIX AT OPTIMAL THRESHOLD- MODEL B

Given an optimal prediction threshold (p_0) of 69.66%, the model correctly predicts the repayment of the loans 98,27% of the times.

7.2.3. Comparison: Before and After the Introduction of the Social Performance Badges

At this point, it is possible to compare all the effects before and after the introduction of the badges with the optimal model from Section 6, where possible. Several variables showed changes in the magnitude of the coefficient estimates across models (see full table in Appendix). **Table 19** compares the variables whose coefficient estimates have also changed sign. These results determine that it is worthwhile to test some interactions effects between loan- and partner-specific predictors, and Social Performance Badges.

Coefficients	Estimates Model 1	Estimates Model A	Estimates Model B
partners.profitability	0,014	-0,004	-0,005
Client_VoiceTRUE	0,181	-	-0,006
score	0,434	0,470	0,130
dummy_Construction	0,100	-0,050	0,000
dummy_Manufacturing	-0,006	0,308	-0,825
dummy_Education	0,120	-0,009	0,000

TABLE 19: COMPARISON OF COEFFICIENT ESTIMATES FROM THREE DIFFERENT MODELS (PARTIAL)

Given the limitations from the small sample size used for Model B, few comments can be drawn from the comparison. The notable differerences among coefficients that were not set to 0 by the lasso regression are those for partners' profitability, sentiment score and for projects in the three sectors. In the first case, partners' profitability, the coefficient estimate in the general model is positive, whereas, in the two time-constrained datasets, it is slightly negative. Instead, projects in the Manufacturing sector are assigned a negative coefficient estimate in Model 1 and in Model B, meaning that, in general and after the introduction of the badges, a project in the "Manufacturing" category would lower the chances of loan repayment. Before the introduction of the badges, however, the coefficient estimate is positive, highlighting a change in the impact of this sector on chances of loan repayment. A similar reasoning applies also to the "Education" and the "Construction" sector. More interestingly, there is a change in the magnitude of the impact of sentiment score on chances of loan success. While in general and before the introduction of the Social Performance Badges the coefficient estimate was between 0,43 and 0,47, after the introduction it dropped to 0,13.

7.3. Implementation of the Findings: Introducing Interaction Effects (Model C)

Given the changes in estimated coefficients before and after the introduction of the Social Performance Badges, a final model is tested, including interaction effects between the presence of the badges and the predictors whose coefficients have undergone a variation (Section 7.2.3.). The new model is built on the full set of data used in sections 5 and 6. All the predictors are used for the prediction purpose. **Table 20** summarizes predictors and interaction terms, divided in the category to which they refer.

Loan-specific Variables	**Partner-specific Variables**	**Sentiment Score**	**Interaction Terms**
loans.sector	partners.status	score	loans.sector*SPB
loans.loan_amount	partners.rating		partners.profitability*SPB
loans.lender_count	partners.delinquency_rate		score*SPB
loans.bonus_credit_eligibility	partners.default_rate		
loans.terms.repayment_interval	partners.total_amount_raised		
loans.terms.repayment_term	partners.loans_posted		
loans.terms.loss_liability.nonpayment	partners.charges_fees_and_interest		
loans.journal_totals.entries	partners.average_loan_size_percent_per_capita_income		
SPB**	partners.loans_at_risk_rate		
	partners.portfolio_yield		
	partners.profitability		

**SPB: the variable measures the presence and number of Social Performance Badges.
min = 0 ; mean = 2.696; max = 7

TABLE 20: LIST OF PREDICTORS AND INTERACTION TERMS - MODEL C

The purpose of this model is to understand how the introduction of the badges has changed the impact of other predictors on the probabilities of loan repayment.

7.3.1. Model C: Discussion of Results

The variables that showed a relevant change in coefficient before or after the introduction of Social Performance Badges, namely the sector of the loan, the sentiment score of text descriptions and the profitability of the Field Partners, are used as interaction terms together with the variable "SPB" ("Social Performance Badges"). This newly introduced variable simplifies the analysis, by measuring the aggregate presence and number of badges per individual funding project. The results of the new model (Model C) are presented in **Table 21**.

Coefficients	**Estimate**	**Std. Error**	**z value**	**A.E.O.**
(Intercept)	16,837	6522,639	0,003	2E+07
factor(loans.sector)Arts	-0,787	0,777	-1,013	0,455
factor(loans.sector)Clothing	-1,599	0,394	-4,055	0,202
factor(loans.sector)Construction	-0,367	0,839	-0,437	0,693
factor(loans.sector)Education	-2,016	1,227	-1,642	0,133
factor(loans.sector)Entertainment	3,049	4,092	0,745	21,104
factor(loans.sector)Food	-1,370	0,319	-4,295	0,254
factor(loans.sector)Health	-1,573	0,801	-1,963	0,207
factor(loans.sector)Housing	-0,279	0,877	-0,318	0,757
factor(loans.sector)Manufacturing	-1,824	1,153	-1,583	0,161
factor(loans.sector)Personal Use	1,418	2,108	0,673	4,128
factor(loans.sector)Retail	-1,614	0,339	-4,760	0,199
factor(loans.sector)Services	-1,172	0,421	-2,781	0,310
factor(loans.sector)Transportation	-1,056	0,688	-1,534	0,348
factor(loans.sector)Wholesale	-2,661	1,047	-2,541	0,070
loans.loan_amount	0,001	0,000	3,214	1,001
loans.lender_count	-0,045	0,014	-3,312	0,956
loans.bonus_credit_eligibilityTRUE	-1,065	0,286	-3,731	0,345
loans.terms.repayment_intervalIrregularly	0,668	0,518	1,291	1,951
loans.terms.repayment_intervalMonthly	0,612	0,400	1,531	1,845
loans.terms.repayment_term	-0,126	0,015	-8,426	0,882
loans.terms.loss_liability.nonpaymentpartner	2,063	0,194	10,655	7,866
loans.journal_totals.entries	-0,441	0,067	-6,601	0,643
partners.statusclosed	0,519	0,631	0,823	1,681
partners.statusinactive	14,255	375,487	0,038	2E+06
partners.statuspaused	-0,269	0,643	-0,418	0,764
partners.rating	0,892	0,224	3,975	2,440
partners.delinquency_rate	-0,031	0,011	-2,811	0,970
partners.default_rate	-0,205	0,011	-18,431	0,815
partners.total_amount_raised	-0,000	0,000	-0,947	1,000
partners.loans_posted	0,000	0,000	2,637	1,000
partners.charges_fees_and_interestTRUE	-11,730	6522,639	-0,002	0,000
partners.average_loan_size_percent_per_capita_income	-0,009	0,001	-6,929	0,991
partners.loans_at_risk_rate	0,028	0,010	2,800	1,028
partners.portfolio_yield	-0,004	0,007	-0,549	0,996
partners.profitability	0,079	0,030	2,674	1,082
SPB	-0,294	0,093	-3,168	0,746
score	0,591	0,055	10,840	1,807
factor(loans.sector)Arts:SPB	0,331	0,299	1,107	1,392
factor(loans.sector)Clothing:SPB	0,277	0,140	1,976	1,319
factor(loans.sector)Construction:SPB	0,170	0,284	0,596	1,185
factor(loans.sector)Education:SPB	1,041	0,548	1,901	2,833
factor(loans.sector)Entertainment:SPB	-1,171	0,889	-1,317	0,310
factor(loans.sector)Food:SPB	0,415	0,116	3,571	1,514

Coefficients	Estimate	Std. Error	z value	A.E.O.
factor(loans.sector)Health:SPB	0,614	0,448	1,369	1,848
factor(loans.sector)Housing:SPB	0,339	0,351	0,967	1,404
factor(loans.sector)Manufacturing:SPB	0,711	0,467	1,522	2,036
factor(loans.sector)Personal Use:SPB	-0,642	0,614	-1,045	0,526
factor(loans.sector)Retail:SPB	0,456	0,124	3,690	1,578
factor(loans.sector)Services:SPB	0,357	0,154	2,315	1,429
factor(loans.sector)Transportation:SPB	0,486	0,266	1,827	1,626
factor(loans.sector)Wholesale:SPB	11,326	410,338	0,028	8E+04
partners.profitability:SPB	-0,022	0,007	-3,104	0,978
SPB:score	-0,053	0,018	-2,928	0,948

A.E.O.: Average Effect on the Odds of success = e^β

Chi-square: 1833.78

Df: 53

TABLE 21: RESULTS FROM LOGISTIC REGRESSION (MODEL C)

From the results, it is possible to draw some considerations. In first place, the aggregate presence of Social Performance Badges has an overall negative impact on the chances of loan repayment. An increase of 1 unit in the aggregate presence of badges leads to a value of 0,74 for the odds of loan success - that is, repayment would only be 3/4 times as likely as default. More interesting results come from the interaction coefficients. For the majority of loan sectors (12 over 14) the presence and number of Social Performance Badges has a positive effect on the chances of loan success. Only for the "Entertainment" and "Personal Use" sectors, an increase in the number of badges has a negative impact on the probability of repayment (the coefficient estimates are -1,17 and -0,64 respectively). For what concerns profitability, the interaction between the number of badges and the profitability of the Field Partner leads to a negative coefficient estimate. Similarly and more unexpectedly, also the interaction between sentiment score and number of badges leads to a decrease in the log odds of loan repayment of -0,053 units.

7.3.2. Model C: Accuracy

The results from a 10-fold cross validation (**Table 22**) show that, at an optimal threshold p_0 of 54,54%, the model predictions for loan repayment are correct 97,99% of the times.

	Glm Predictions*	
Observed Outcome	**Defaulted**	**Paid**
Defaulted	177	62
Paid	237	14401

*p_0 = 54.54% , n = 14877

TABLE 22: CONFUSION MATRIX AT OPTIMAL THRESHOLD- MODEL C

7.4. General Discussion

7.4.1. Situation Before December 11st, 2011

From the analysis of the observations posted before the introduction of the Social Performance Badges, it appears that findings from Kumar (2007) and Everett (2015) about the impact of loan amount on chances of repayment are not confirmed. In the sample in use, an increase in loan amount has zero or small positive effect on probabilities of loan success. On contrary, findings from Dorfleitner et al. (2016) and Gao and Lin, (2015) are confirmed, with positive text descriptions having the biggest impact on loan repayment probabilities when compared to the overall model (Model 1) and the model with more recent data (Model B). This also confirms the research expectations for the present research. For what concerns the expectations about Social Performance Badges, it is not possible to draw conclusions in that respect, since they were not present at the time the observations were recorded.

Additional insights from the analysis show that the biggest positive contribution to loan success came from the status of the Field Partner (which is, however, biased by the possible change in status over time) and the attribution of liability in case of losses. Projects in the Manufacturing sector appeared to have a more positive impact on loan success in the past when compared to more recent times - while the contrary is true for projects in the Education sector. Among loan-specific factors, repayment terms have impact positively the chances of loan repayment, contrary to what found by Field and Pande (2008). Partner-specific factors that have positive effect on loan success are instead the rating and, unexpectedly, the loan-at-risk rate — although, for the latter, the positive coefficient is very close to 0.

7.4.2. Situation After December 11st, 2011

After the introduction of Social Performance Badges, the loan amount's coefficient estimate decreased and was set to 0, having no effect on the probability of loan success. This somewhat confirms the findings from Kumar (2007) and Everett (2015), although any statement in this sense could be biased, given the small sample size used for the analysis. Findings from Dorfleitner et al. (2016) and Gao and Lin, (2015) are once more confirmed, reinforcing the expectation that text description can, in fact, successfully play a role in predicting project success also on other platforms. All Social Performance Badges, except for the Client Voice badge, seemed to have no clear effect on the probability of loan repayment in the subset of 580 observations. This may be due to the fact that the vast majority of the observations had a posted date between December 12nd, 2011 and February 3rd, 2012, meaning that the lenders had little time to get acquainted to the use of these symbols. The number of lenders has a negative impact on the chances of loan repayment, together with an increase in repayment term. While the first is unexpected, the second appears to be reasonable in such a micro-finance context. A higher number of journal entries also seems to damage the outcomes of the project — maybe because risky or uncertain loans require more frequent updates. In line with expectations, the higher the delinquency and default rates of Field Partners, the higher the chances of loan default. An increase in the average loan size as a percentage of per capita income is also associated with loans less likely to be repayed. Increases in portfolio yields of partners are also liked to loans more likely to default - probably because the more risky the project, the higher the interests and fees charged to the borrower of the loan, the more likely the project is to default within the established time limit.

7.4.3. Discussion of Results: Sentiment Score and Social Performance Badges

Both the general model (previously addressed as Model 1) and the model with interactions (Model C) confirmed all the research expectations expressed early in the present research. In particular, they confirmed that the valence of text descriptions can be effective predictors of chances of loan repayment, as suggested by Gao and Lin, (2015). This also confirms that the findings from Gao and Lin, (2015) are suitable to generalization to platforms other than Prosper.com. Also, the expectation that both the presence and the number of Social Performance Badges play a role on the probabilities of loan repayment is confirmed. The two effects were studied separately, both as main effects and, for what concerns the number of badges, also as an interaction term. None of the effects of the badges was penalized by the lasso regression of Model 1. Nearly half of them (Antipovery, Vulnerable Group and Innovation) have actually a negative impact on chances of loan repayment (see results from Model 1, Section 6). Moreover, the more badges, the less chances for a loan to be paid back (see Model C). This probably happens because of the lack of specialization of the managing partner, although the subject matter has not yet been researched, and any statement in this respect would require further investigation. On the other hand, the remaining half of the badges have a positive impact on the probabilities of loan success, leading to an overall satisfactory research outcome.

7.4.4. Discussion of Results: Loan Specific Variables

On aggregate, loan-specific factors that increase the odds of loan success are the repayment interval (irregular or monthly, as opposed to "end of the term") and the liability in case of losses attributed to the Field Partner. This may be because more frequent repayments and partners willing to take the loss liabilities might suggest financially "healthy" projects. On contrary, increases in repayment terms and journal entries, as well as backers to the loan and the possibility for a bonus credit do not contribute positively to the chances of loan success. As discussed before, unsuccessful loans may require more frequent updates, more backers and, in general, more time to be paid back.

7.4.5. Discussion of Results: Partner Specific Variables

As expected, partner's default and delinquency rates, together with portfolio yields, have negative effects on probabilities of repayment. The contrary happens for increases in rating. Eventually, while higher loan-at-risk rate are associated with higher chances of loan repayment, a higher average loan size as percent of per capita income is generally related to projects more likely to default. The first result is unexpected, since it would be reasonable to expect higher amounts of loan-at-risk to be associated with projects more likely to default.

7.4.6. Discussion of Results: Interaction Effects

Interaction effects also confirmed the expectations: the number of Social Performance Badges interacts with both loan sector, partner's profitability and text description sentiment scores. The interactions of badges with loan sectors are various, but generally positive: the presence and number of badges positively interacts with 11 sectors towards the probabilities of loan repayment. This means that, those same projects, within the same sectors, would be associated with less chances to succeed with a lower number of Social Performance Badges. On contrary, the presence of Social Performance Badges is not contributing positively to the project chances to succeed when the Field Partner is more profitable ($\beta_{PROF:SPB}$ = -0,022) or when the text description is more positive ($\beta_{S:SPB}$ = -0,053), putting the interaction effect in line with the negative main effect of the number of badges (β_{SPB} = -0,294).

8. Conclusions

8.1. Overview of the Findings

In the present research, the probabilities of repayment of microloans are studied in the online prosocial context of a non-market organization, Kiva.org. Several models with different characteristics and purposes were trained and tested to assess the impact of certain features on the chances of loan success. While one model (Model 1) had a general approach to the problem, Models A and B were run based on chronological criteria, to understand the changes brought to repayment dynamics by the introduction of Social Performance Badges. The last model (Model C) implemented the insights learned from the three models, introducing interaction terms into the predictive attempt. Broadly speaking, it is possible to notice a certain consistency in the findings of the four models. The findings are summarized in **Table 23**.

	Model 1	**Model A**	**Model B**	**Model C**
Description	Full dataset after variable selection	Before the introduction of the badges, full list of predictors minus SPBs	After the introduction of the badges, predictors subject to shrinkage	Full datset, implement insights from previous models by introducing interaction effects
Factors Increasing Chances of Repayment	Sentiment score, 4/6 SP badges, partner's loss liability non-payment, loan amount, partner rating, partners' loan at risk rate, partner status, partner profitability, 5 loan sectors, repayment interval.	Sentiment score, 4 loan sectors, partner's loss liability non-payment, partner status, partner rating, partners' loan at risk rate.	Sentiment score.	Sentiment score, loan amount, repayment interval, partner's loss liability non-payment, partner status, partner rating, partner loan at risk rate, partner profitability.
Factors Decreasing Chances of Repayment	Repayment term, partner delinquency and default rate, partner average loan size as percent of per capita income, 2/6 SP badges, lender count, bonus credit eligibility, journal entries, partner portfolio yield, partner charges fees and interest.	Repayment term, partner delinquency and default rate, partner average loan size as percent of per capita income, lender count, bonus credit eligibility, journal entries, partner portfolio yield, partner profitability.	Repayment term, partner delinquency and default rate, partner average loan size as percent of per capita income, 2 loan sectors, lender count, journal entries, partner portfolio yield, partner profitability, Client Voice SP badge.	Repayment term, number of SP badges, Entertainment:SPB, Personal Use:SPB, partner profitability:SPB, sentiment score:SPB, 12 loan sectors, lender count, bonus credit eligibility, journal entries, partner delinquency and default rate, partner average loan size as percent of per capita income, partner charges fees and interest, partner portfolio yield.
Nr. of Obs.	14877	14281	580	14877
Nr. of Predictors	41	35	46	53
Prediction Method	Lasso Regression	Logistic Regression	Lasso Regression	Logistic Regression
Prediction Accuracy	97,97%	97,97%	98,27%	97,99%

TABLE 23: SUMMARY OF THE RESEARCH FINDINGS

8.1.1. Social Performance Badges and Sentiment Score

The research expectations are always confirmed for what concerns the sentiment score, which has been found to have a positive effect on probabilities of loan repayment in all models. This is also in line with the findings of Dorfleitner et al. (2016) and Gao and Lin (2015).

Contrasting effects were instead observed for the presence of individual Social Performance Badges. Only in the first model (Model 1), the majority of the badges, namely Client Voice, Family and Community, Enterpreneurial Support and Facilitation of Savings, had a positive impact on chances of repayment. In Model B, the only badge not shrinked to zero was the Client Voice badge, which, however, was associate with a negative effect. For what concerns the aggregate number of badges, it was associated with negative impact on chances of repayment. However, in interaction with the different loan sectors, it turned out to positively affect probabilities of success. This finding may suggest that certain field partners (those granted with more badges and likely to operate in a less specialized manner) may be more successful at operating in certain sectors rather than others. For example, projects in the "Education" sector have shown higher probabilities of loan repayment when associated to a higher number of Social Performance Badges. Profitability and sentiment score have a negative impact on loan success when interacting with the number of Social Performance Badges. This suggests that more specialization (in the form of fewer badges), together, in turn, with more profitable partners and positive text descriptions, are on average associated with higher chances of successful loan repayment.

8.1.2. Other Predictors

For what concerns loan- and partner-specific predictors, consistent results are found about the positive impact of partner's loss liability non-payment, loan amount, partners' rating, partners' loan at risk rate and partner profitability. Certain levels of partners' status have apparently positive estimated coefficients, although several limitations exist for this variable. The partner's loss liability non-payment may be a signal that a partner is confident that the loan won't default, and hence is ready to take the liability in case loss from non-repayment should occur. Bigger loan amounts may be instead more difficult to handle or repay timely by the borrowers, and thus lead to higher chances of default. Partners that were rated more positively and with higher profitability figures may have proven themselves more reliable than others, which reinforced their association with higher chances of repayment. An uncertain finding is the effect of the loan-at-risk rate, which would be expected to lower chances of repayment and is instead consistently associated with higher probabilities of success.

Consistently negative factors are repayment term, partner delinquency and default rate, partner average loan size as percent of per capita income, lender count, bonus credit eligibility, journal entries and partner portfolio yield. As already discussed earlier, longer repayment terms may signal borrower's problems in the repayment process, leading to higher chances of loan default. Similarly, problematic loans may have necessity for more journal updates and are associated with higher interests and yields by the field partners. Field partners may find more difficult to administer successfully larger average sizes of their loans. Similarly, problematic loans may require more lenders to be founded when compared to successful, non-problematic projects.

8.2. Implications of the Research Findings for Practitioners on Kiva

The findings of this study partially confirmed the results of previous research, complementing them with additional, platform-specific details. Parties involved in transactions on Kiva can definitely make use of the additional knowledge derived from the present research. Among the many models tested, practitioners can certainly use Model 1 for prediction purposes: the model is complete with all the relevant predictors while using 12 less variables than Model C, and achieves a satisfying 97,97% prediction accuracy. Model C was, in fact, developed to understand the impact of the Social Performance Badges on variables that showed a change before and after their introduction. For prediction purposes, however, a simpler model performs just as well, while using fewer predictors. In Appendix, an alternative, non-linear prediction model is developed as an additional instrument for practitioners' use.

8.2.1. Field Partners

Field Partners can reason about whether it is worth or not to obtain and add one or more accreditation badges to their profiles. In certain cases - especially if they operate in determined sectors - it may be worth to be specialized in one or few social missions rather than aiming at too many social performance accreditations. For example, partners working in the Entertainment and Personal Use sectors may want to be more specialized, since the interaction between the number of badges and these sectors causes has a potential for unsuccessful projects. Besides the particular case of the accreditation system, Field Partners can find in the predictive models useful tools to individuate potentially hazardous situations. In turn, this may also prevent the occurrence of the dangerous, downward spiral of over-indebtness.

8.2.2. Lenders

Kiva lenders can make use of the general findings from this study to have a better, preliminary understanding of the success potential of their lending portfolio. As mentioned in Seciton 2, in fact, lenders can improve loan selection by collecting private information about the borrowers (Fama, 1985). Recurring factors in the positive or negative spectrum of loan success may be particularly useful for their assessment.

8.2.3. Kiva Operators

Kiva operators and associates can make use of the findings about text descriptions to better market potentially successful projects. Notice that, since correlation is not equal to causation, crafting a positive text description does not ensure the success of a project. Rater, a well-marketed descritpion can be a powerful tool to enhance the pre-existing strengths of the project at stake. Kiva can also use the models to monitor the projects and spot potentially hazardous situations. These would be situations in which default is very likely to occur and timely action can prevent the occurrence of unpleasant events, like over-indebtness.

8.3. Limitations to the Present Research

There are several limitations to the present research, the majority of which can be traced back to a small sample size. Initially, a sufficiently large sample was selected, which was however reduced considerably after all the constraints applied in the choice of the dependent and independent variables. Further research is encouraged, to repeat the findings with a larger sample size. Suggested sizes may be the double or triple of the size used in the present research.

Another limitation comes from the Social Performance Badge variables. First of all, they are a relatively new addition to the features of Kiva website, which might mean that Kiva lenders are not yet entirely familiar with their meaning and usage. Moreover, in the dataset selected, the majority of the observations were not posted after 2013, representing a definitely too short time span for results to be considered satisfactory. This limitation is confirmed by the fact that the lasso regression for Model B shrinked all the SP Badges coefficients, but one, to zero. Again, a larger sample size with more up-to-date observations, possibly collected later in time than the present date, would easily solve this limitation. The consequences of these limits are to be seen in the results of the analyses. Additionally, as already specified in Section 4, the Badges are assigned yearly based on a process of due diligence. Essentially, this means that the Badge variables are likely to change over time, if a partner does not meet the requirements during the due diligence process. However, in the dataset and during the analysis, badges are considered as static elements. As this is not the case in reality, the assumption represents a big limitation. Should futher research take place in this setting, a time-based comparison of Social Performance Badges dynamics is encouraged. For what concerns the model accuracy, the number of false positives and false negatives is relatively high, considering that the objective of the models is the detection of hazardous situations. Also, some variables present singularities and had to be excluded from the picture, and a logistic regression applied to Model B would entail a perfect prediction split, with increase in inaccuracy and bias. Eventually, a limitation comes from the partner status variable. In fact, the status of a Field Partner may have changed over time. This is a problem similar to the Social Performance variables: the associated results are not corrected for this bias and, as a consequence, must be interpreted with due care.

8.4. Suggestions for Further Research

In spite of the limitations, the present research led to a great amount of interesting suggestions for future research. Since the accreditation system proved to have predictive potential, it may be interesting to observe how a similar system would interact with different environments. Lab studies that use accreditation token as control variables can help complement the findings of the present research. The results also highlighted that a higher number of badges for a partner lowers the chances of loan success. Future research can focus on the issue of the lack of specialization of the managing partner and assess whether the finding is grounded on evidence in this sense.

Also, the study of sentiment scores could be extended to text descriptions in other languages. This would require the creation and use of lists of valenced words in languages other than English. A country-of-origin effect may be assessed on the basis of the description language. Since there are no details about the content of the journal entries, it would be worthwhile to investigate the textual elements in light of the findings of Gao and Lin (2015) about the objectivity of project information. It could be interesting to understand whether also the journal updates, together with text descriptions, can be successful predictors of loan repayment. Since the higher number of journal entries also seems to damage the outcomes of the project, future research can investigate whether a case exists for risky or uncertain loans to require more frequent updates. Additionally, since the average loan size as percent of per capita income disbursed by the Field Partner has a negative effect on chances of repayment, research is suggested to understand whether higher amounts at stake, in this context, are actually exposed to additional sources of risk.

References

Ashta, A., & Bumacov, V. (2011). From Social Rating to Seal of Excellence: Utility or Futility?. Available at SSRN 1890682.

Bergeron, JD. (2011, December 11). Kiva Launches Social Performance Badges and Increases the Information Available for Your Lending Decisions [blog post]. Retrieved from https://www.kiva.org/blog/kiva/2011/12/11/kiva-launches-social-performance-badges-and-increases-the-information-available-for-your-lending-decisions.html

Breen, J. O. (2012). Mining twitter for airline consumer sentiment. *Practical text mining and statistical analysis for non-structured text data applications*, **133**.

Bruett, T. (2007). Cows, kiva, and prosper. com: How disintermediation and the internet are changing microfinance. *Community Development Investment Review*, **3**(2), 44-50.

Coleman, R. W. (2007). Is the Future of the Microfinance Movement to be Found on the Internet? *International Trade and Finance Association Conference Papers* (p. **1**). bepress.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending–Evidence from two leading European platforms. *Journal of Banking & Finance*, **64**, 169-187.

Everett, C. R. (2015). Group membership, relationship banking and loan default risk: the case of online social lending. *Banking and Finance Review*, **7**(2).

Fama, E. F. (1985). What's different about banks?. *Journal of monetary economics*, **1**.

Fenton, A. (2010, December 21). The misleading metrics of microcredit [blog post]. Retrieved from http://www.iied.org/misleading-metrics-microcredit.

Field, E., & Pande, R. (2008). Repayment frequency and default in microfinance: evidence from India. *Journal of the European Economic Association*, **6**(2‐3), 501-509.

Galak, J., Small, D., & Stephen, A. T. (2011). Microfinance decision making: A field study of prosocial lending. *Journal of Marketing Research*, **48**(SPL), S130-S137.

Gao, Q., & Lin, M. (2015). Lemon or Cherry? The Value of Texts in Debt Crowdfunding. *The Value of Texts in Debt Crowdfunding* (July 4, 2015).

Gareth, J. (2013). *An Introduction to Statistical Learning: with Applications in R* (p. 146). Springer.

Godquin, M. (2004). Microfinance repayment performance in Bangladesh: How to improve the allocation of loans by MFIs. *World Development,* **32**(11), 1909-1926.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* 168-177.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (**Vol. 112**). New York: springer.

Kumar, S. (2007). Bank of one: Empirical analysis of peer-to-peer financial marketplaces. *AMCIS 2007 Proceedings*, 305.

Ledgerwood, J. (1998). Microfinance handbook: An institutional and financial perspective. *World Bank Publications*.

Lee, E., & Lee, B. (2012). Herding behavior in online P2P lending: An empirical investigation. *Electronic Commerce Research and Applications,* **11**(5), 495-503.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, **183**(3), 1466-1476.

Microfinance Information Exchange Market (2003). Focus on Savings: Microbanking Bulletin: 9th Issue.

R Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria. URL https://www.R-project.org/.

Schicks, J. (2010). Microfinance Over-Indebtedness: Understanding its drivers and challenging the common myths. *Bruxelles: Centre Emilee Bergheim, Solvay School of Business, CEB Working Paper*, **10**, 048.

Web Resources

R codes to replicate analyses performed in this study: https://github.com/elisewinn/Pocchiari_2016_Thesis

General Q&As about Kiva.org: http://blog.kiva.org/faqs/kivas-impact-qa

Alternative platforms similar to Kiva: https://www.quora.com/What-are-some-good-alternatives-to-Kiva

Kiva Field Partners: http://www.kiva.org/partners

APPENDIX

Table 1: Detailed Overview of Loan Data (Dataset 1.json, date 01/01/2016)

Column name	Information contained	Measurement Level
header.total	Total number of items in the directory	Interval/Ratio
header.page	Page of data in the series	Interval/Ratio
header.date	Timestamps referred to the project, according to ISO 8601	Date
header.page_size	Total number of entries per page	Interval/Ratio
loans.id	Integer used to refer to an object	Interval/Ratio
loans.name	First name of the borrower	Categorical
loans.description	The description field has different sub-categories	Categorical
loans.description.languages	Languages of the text description	Categorical
loans.description.texts.en/es/fr/ru	Text description of the project in English, Spanish, French or Russian	Categorical
loans.status	• Fundraising: the loan has not yet been funded. • Funded: this loan request has been completely funded and is not available for new loans by lenders. • In_repayment: the loan has been disbursed to the borrowers and they are in the process of using the funds and making payments on the loan to the field partner. • Paid: the loan has been paid back in full by the borrower. The payments have been distributed back to the lenders and the loan is closed to most new activity on Kiva. • Defaulted: a borrower or a field partner fails to make payments on a loan. When this happens, a loan becomes delinquent. When a loan remains delinquent 6 months after the end of the loan payment schedule, the loan becomes defaulted	Categorical
loans.funded_amount	Amount funded so far	Interval/Ratio
loans.basket_amount	Amount of the loan which lenders have saved in shopping baskets, but has not been confirmed as purchased. It is possible that in the near future a portion or all of this amount could become available to other lenders, but until that time this amount of the loan is reserved	Interval/Ratio
loans.paid_amount	Amount paid back by the borrower	Interval/Ratio
loans.image.id and loans.image.template_id	An image in the Kiva API is represented by an id. The specific template you can use for an image can depend on the photo	Interval/Ratio
loans.video	Video in loan profiles and journal entries	NA
loans.activity	Activity that will be performed with the funded amount	Categorical

Column name	Information contained	Measurement Level
loans.sector	Industrial sector to which the activity belongs	Categorical
loans.themes	Attributes of the project	Categorical
loans.use	Brief text description of the use that will be made of the funded capital	Categorical
loans.delinquent	Whether the loan is delinquent or not (admits TRUE, FALSE values)	Logical (True; False)
loans.location.country_code	2-letter code of the country of origin of the borrower	Categorical
loans.location.country	Full name of the COO	Categorical
loans.location.town	Town of origin of the borrower	Categorical
loans.location.geo.level	Reflects the level of accuracy available for the supplied geometry. Popular values might be town, country, or exact. Use this information to help you decide how to represent the data, or if you should do your own geocoding for the object.	Categorical
loans.location.geo.pairs	The coordinate pairs for the geometry. This value is formatted according to the GeoRSS-Simple specification for serializing coordinates.	Interval/Ratio (Pair)
loans.location.geo.type	The type of geometry defined by the coordinate pairs provided. This can be any of the shapes supported by the GeoRSS Model — point, line, box, or polygon.	Categorical
loans.partner_id	Field Partner ID	Interval/Ratio
loans.posted_date	Date the loan was posted	Date
loans.planned_expiration_date	Date in which the loan is planned to expire	Date
loans.loan_amount	Amount asked by the borrower	Interval/Ratio
loans.lender_count	Number of lenders to the project	Interval/Ratio
loans.currency_exchange_loss_amount	If a currency exchange loss is shared between a partner and the lender, then this value represents the amount in USD lost by the lender due to fluctuations in the value of the local currency against the US dollar.	Interval/Ratio
loans.bonus_credit_eligibility	A 25$ bonus per friend invited. Admits values TRUE or FALSE	Logical (True; False)
loans.tags	Tags assigned to the loan project	Categorical
loans.borrowers	Information about the borrower(s)	Categorical
loans.terms.disbursal_date	The date at which the funds from the loan were given to the borrowers.	Date
loans.terms.disbursal_currency	The ISO 4217 code for the currency used to distribute the loan the the borrower. This is usually the local currency for the borrower's country.	Categorical

Column name	Information contained	Measurement Level
loans.terms.disbursal_amount	The amount of money distributed to the borrower(s) in the local currency.	Interval/Ratio
loans.terms.repayment_interval	Scheduled rate of repayment (monthly, at the end of the term…)	Categorical
loans.terms.repayment_term	Term set for the repayment of the loan	Interval/Ratio
loans.terms.loan_amount	View loans.loan_amount	Interval/Ratio
loans.terms.local_payments	Payments disbursed by borrower	Date
loans.terms.scheduled_payments	Payments scheduled in future time periods	Date
loans.terms.loss_liability.	A set of values which describes who is liable for loss on a loan, and how. For nonpayment, the party liable can either be the lender or partner. For currency_exchange the liability can be shared, fully resting on the partner, or none if the currency is locked to the US dollar. If currency exchange loss is a shared liability, the currency_exchange_coverage_rate will also be listed.	Categorical
loans.payments	Admits list() value	list()
loans.funded_date	Date the loan was fully funded by lenders	Date
loans.paid_date	Date the loan was fully repaid by borrower	Date
loans.journal_totals	The number of total journal entries for the loan as well as the number which are automated or "bulk" entries. Bulk entries are typically much less interesting than other journal entries. Checking these counts can help you determine if it is worthwhile for your application to fetch the journals for a loan.	Interval/Ratio
loans.translator.byline	Name of the translator of texts	Categorical
loans.translator.image	Image of the text translator	Categorical

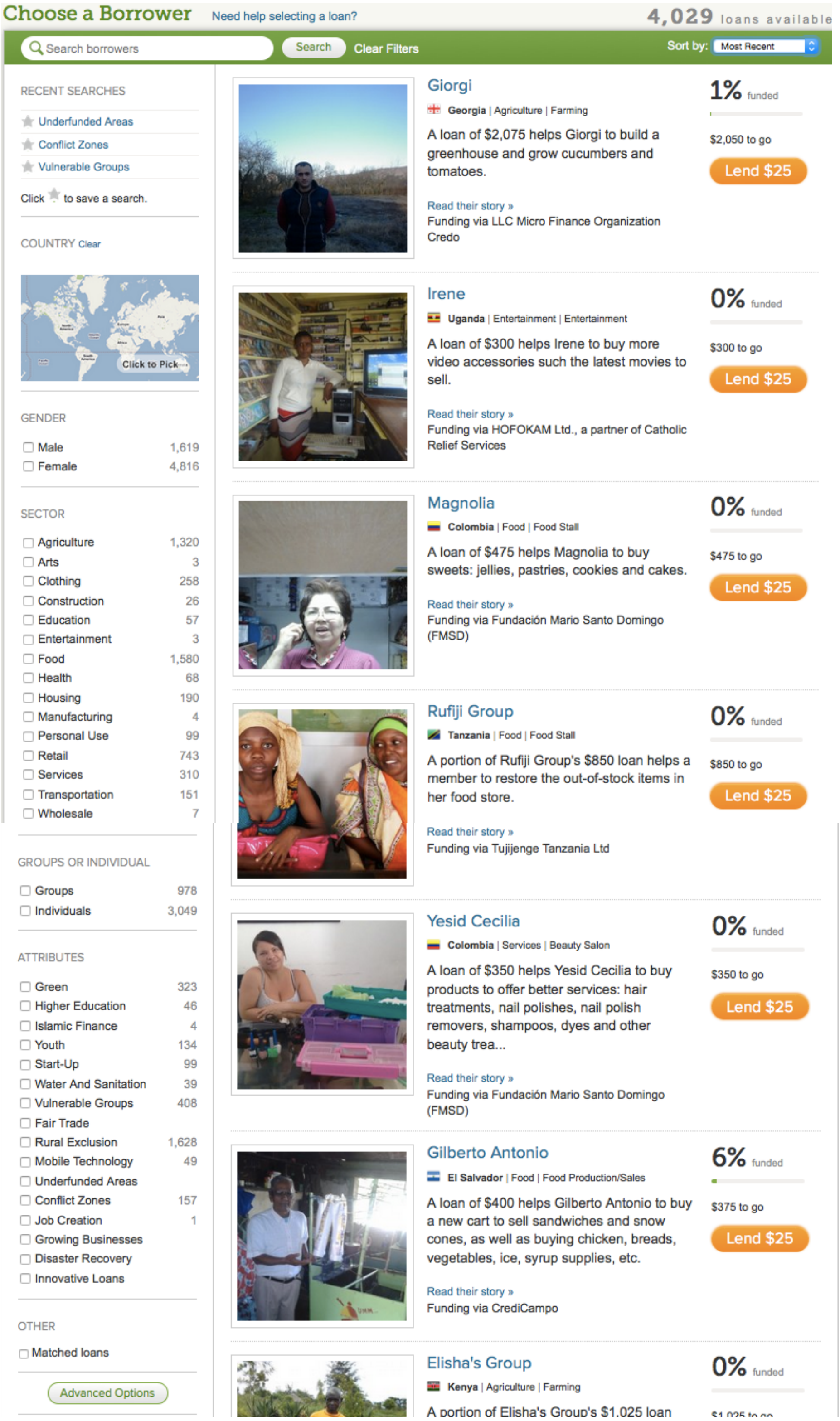
Table 2: Detailed Overview of Partners' Data (Dataset partners.json, date 01/01/2016)

Column name	Information contained	Measurement Level
paging.page, paging.total, paging.page_size, paging.pages	Information about the web pages containing Partners' information	Interval/Ratio
partners.id	Unique identification id number of the Field Partner	Interval/Ratio
partners.name	Name of the field partner	Categorical
partners.status	whether or not a Field Partner is actively raising funds on Kiva's platform. Kiva adjusts the fundraising status of each Field Partner based on time and performance. Combined with the Field Partner risk rating, the fundraising status determines the number of loans a Field Partner can post in a given month. Kiva Field Partners can be in any of the following statuses: Active: An active partner is currently raising funds on Kiva. Inactive: An inactive partner is not currently posting new loans, but is in good standing with Kiva. A partner may be in inactive status for a variety of reasons, such as operational constraints preventing them from posting loans. Kiva monitors inactive partners according to their outstanding balance, with a higher level of monitoring conducted on partners with higher balances. Paused: A paused partner has been restricted from raising funds while Kiva reviews a potential issue. A partner may be paused for a variety of reasons, such as excessive delinquency or default on Kiva loans, a violation of Kiva policies, or external factors such as country-related regulatory constraints. Closed: A partnership will be closed when Kiva no longer works with this organization.	Categorical
partners.rating	The risk rating (0.5 - 5 stars) reflects the risk of institutional default associated with each of Kiva's Field Partners. Admits "Not Rated"	Categorical
partners.image.id and image.template.id	An image in the Kiva API is represented by an id. The specific template you can use for an image can depend on the photo	Interval/Ratio
partners.start_date	Date of start of the partner's activity	Date
partners.countries	Country of origin of the field partner	Dataframe List
partners.delinquency_rate	Kiva defines the Delinquency (Arrears) Rate as the amount of late payments divided by the total outstanding principal balance Kiva has with the Field Partner. Arrears can result from late repayments from Kiva borrowers as well as delayed payments from the Field Partner. Delinquency (Arrears) Rate = Amount of Paying Back Loans Delinquent / Amount Outstanding	Interval/Ratio

Column name	Information contained	Measurement Level
partners.default_rate	The default rate is the percentage of ended loans (no longer paying back) which have failed to repay (measured in dollar volume, not units). Default Rate = Amount of Ended Loans Defaulted / Amount of Ended Loans	Interval/Ratio
partners.total_amount_raised	Total amount raised by the field partner	Interval/Ratio
partners.loans_posted	Number of loans backed by the field partner	Interval/Ratio
partners.delinquency_rate_note; partners.default_rate_note; partners.portfolio_yield_note;	Notes on the rates of delinquency, default and yield of the field partner	Categorical
partners.charges_fees_and_interest	Admits TRUE, FALSE entries	Logic (TRUE, FALSE)
partners.average_loan_size_percent_per_capita_income	A Field Partner's average loan size is expressed as a percentage of the country's gross national annual income per capita. Loans that are smaller (that is, as a lower percentage of gross national income per capita) are generally made to more economically disadvantaged populations. However, these same loans are generally more costly for the Field Partner to originate, disburse and collect.	Interval/Ratio
partners.loans_at_risk_rate	The loans at risk rate refers to the percentage of Kiva loans being paid back by this Field Partner that are past due in repayment by at least 1 day. This delinquency can be due to either non-payment by Kiva borrowers or non-payment by the Field Partner itself. Loans at Risk Rate = Amount of paying back loans that are past due / Total amount of Kiva loans outstanding	Interval/Ratio
partners.currency_exchange_loss_rate	Amount of Currency Exchange Loss / Total Loans.	Interval/Ratio
partners.url	Field Partner's website, if available.	Categorical
partners.portfolio_yield	a Field Partner's financial earnings divided by its average loan portfolio outstanding during a given year.	Interval/Ratio
partners.profitability	An indication of a Field Partner's profitability. It can also be an indicator of the long-term sustainability of an organization, as organizations consistently operating at a loss (those that have a negative return on assets) may not be able to sustain their operations over time.	Interval/Ratio
partners.social_performance_strengths	Social performance badges granted to the field partner	Dataframe List

Figure 1: Kiva Website Interface and Loan Initial Page

Figure 1 shows the webpage that users face when browsing for loans, with results sorted by most recent. On the next page, a screenshot from the page of an individual open loan is reported. The URL of the webpage is http://www.kiva.org/lend?sortBy=newest, visited on date January, 19th 2016. The URL of the loan page is https://www.kiva.org/lend/1003515, visited on January 31st, 2016.



Choose a Borrower   Need help selecting a loan?   4,029 loans available

Search borrowers   Search   Clear Filters   Sort by: Most Recent

RECENT SEARCHES
- Underfunded Areas
- Conflict Zones
- Vulnerable Groups

Click to save a search.

COUNTRY Clear

GENDER	
□ Male	1,619
□ Female	4,816

SECTOR	
□ Agriculture	1,320
□ Arts	3
□ Clothing	258
□ Construction	26
□ Education	57
□ Entertainment	3
□ Food	1,580
□ Health	68
□ Housing	190
□ Manufacturing	4
□ Personal Use	99
□ Retail	743
□ Services	310
□ Transportation	151
□ Wholesale	7

GROUPS OR INDIVIDUAL	
□ Groups	978
□ Individuals	3,049

ATTRIBUTES	
□ Green	323
□ Higher Education	46
□ Islamic Finance	4
□ Youth	134
□ Start-Up	99
□ Water And Sanitation	39
□ Vulnerable Groups	408
□ Fair Trade	
□ Rural Exclusion	1,628
□ Mobile Technology	49
□ Underfunded Areas	
□ Conflict Zones	157
□ Job Creation	1
□ Growing Businesses	
□ Disaster Recovery	
□ Innovative Loans	

OTHER

□ Matched loans

Advanced Options

Giorgi
Georgia | Agriculture | Farming
A loan of $2,075 helps Giorgi to build a greenhouse and grow cucumbers and tomatoes.
Read their story »
Funding via LLC Micro Finance Organization Credo
1% funded
$2,050 to go
Lend $25

Irene
Uganda | Entertainment | Entertainment
A loan of $300 helps Irene to buy more video accessories such the latest movies to sell.
Read their story »
Funding via HOFOKAM Ltd., a partner of Catholic Relief Services
0% funded
$300 to go
Lend $25

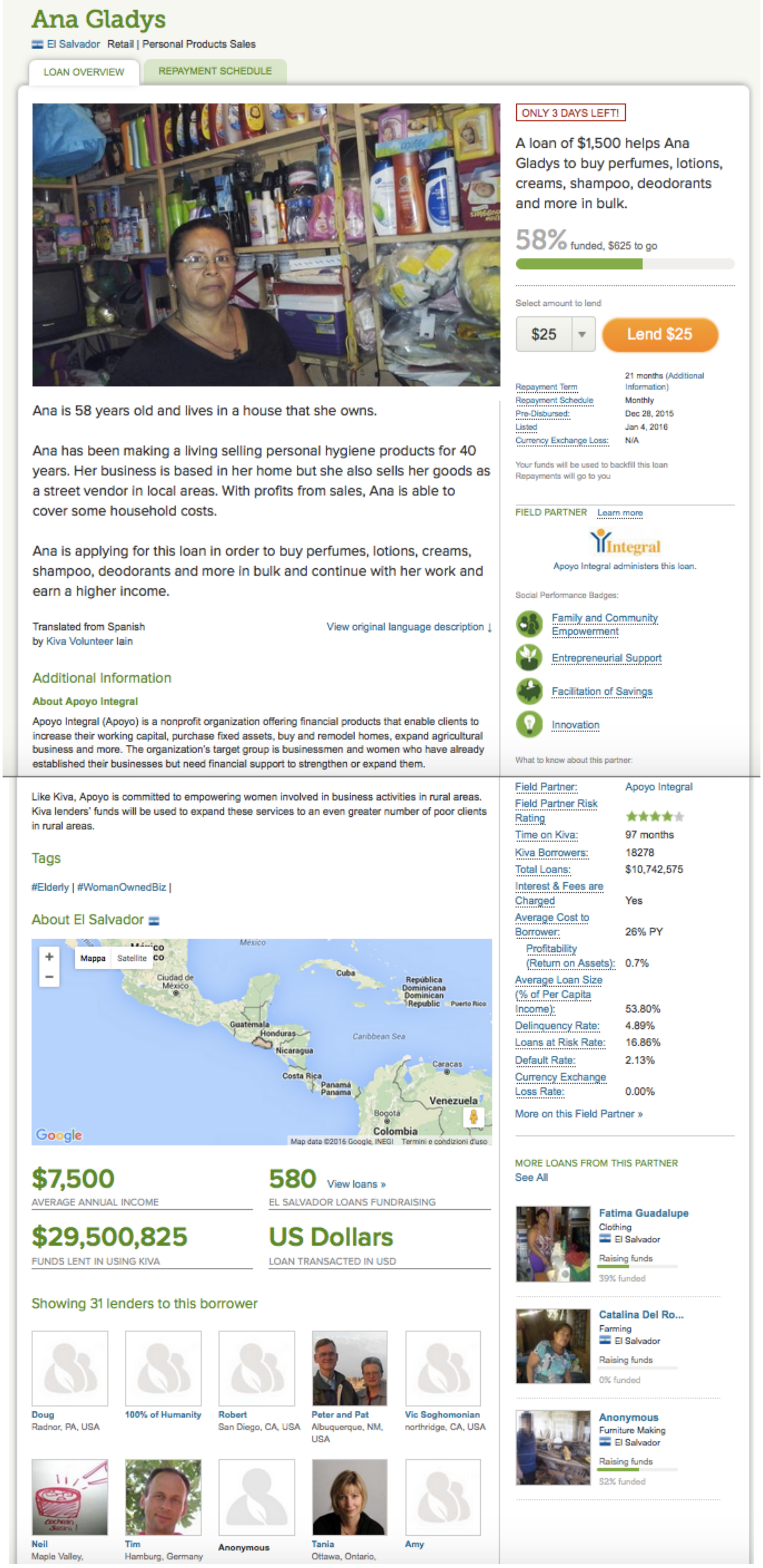
Magnolia
Colombia | Food | Food Stall
A loan of $475 helps Magnolia to buy sweets: jellies, pastries, cookies and cakes.
Read their story »
Funding via Fundación Mario Santo Domingo (FMSD)
0% funded
$475 to go
Lend $25

Rufiji Group
Tanzania | Food | Food Stall
A portion of Rufiji Group's $850 loan helps a member to restore the out-of-stock items in her food store.
Read their story »
Funding via Tujijenge Tanzania Ltd
0% funded
$850 to go
Lend $25

Yesid Cecilia
Colombia | Services | Beauty Salon
A loan of $350 helps Yesid Cecilia to buy products to offer better services: hair treatments, nail polishes, nail polish removers, shampoos, dyes and other beauty trea...
Read their story »
Funding via Fundación Mario Santo Domingo (FMSD)
0% funded
$350 to go
Lend $25

Gilberto Antonio
El Salvador | Food | Food Production/Sales
A loan of $400 helps Gilberto Antonio to buy a new cart to sell sandwiches and snow cones, as well as buying chicken, breads, vegetables, ice, syrup supplies, etc.
Read their story »
Funding via CrediCampo
6% funded
$375 to go
Lend $25

Elisha's Group
Kenya | Agriculture | Farming
A portion of Elisha's Group's $1,025 loan
0% funded
$1,025 to go



Ana Gladys

El Salvador Retail | Personal Products Sales

LOAN OVERVIEW REPAYMENT SCHEDULE

ONLY 3 DAYS LEFT!

A loan of $1,500 helps Ana Gladys to buy perfumes, lotions, creams, shampoo, deodorants and more in bulk.

58% funded, $625 to go

Select amount to lend

$25 ▾ Lend $25

Repayment Term	21 months (Additional Information)
Repayment Schedule	Monthly
Pre-Disbursed:	Dec 28, 2015
Listed	Jan 4, 2016
Currency Exchange Loss:	N/A

Your funds will be used to backfill this loan
Repayments will go to you

Ana is 58 years old and lives in a house that she owns.

Ana has been making a living selling personal hygiene products for 40 years. Her business is based in her home but she also sells her goods as a street vendor in local areas. With profits from sales, Ana is able to cover some household costs.

Ana is applying for this loan in order to buy perfumes, lotions, creams, shampoo, deodorants and more in bulk and continue with her work and earn a higher income.

Translated from Spanish
by Kiva Volunteer Iain

View original language description ↓

Additional Information

About Apoyo Integral

Apoyo Integral (Apoyo) is a nonprofit organization offering financial products that enable clients to increase their working capital, purchase fixed assets, buy and remodel homes, expand agricultural business and more. The organization's target group is businessmen and women who have already established their businesses but need financial support to strengthen or expand them.

Like Kiva, Apoyo is committed to empowering women involved in business activities in rural areas. Kiva lenders' funds will be used to expand these services to an even greater number of poor clients in rural areas.

Tags

#Elderly | #WomanOwnedBiz |

About El Salvador

$7,500
AVERAGE ANNUAL INCOME

580 View loans »
EL SALVADOR LOANS FUNDRAISING

$29,500,825
FUNDS LENT IN USING KIVA

US Dollars
LOAN TRANSACTED IN USD

Showing 31 lenders to this borrower

Doug
Radnor, PA, USA

100% of Humanity

Robert
San Diego, CA, USA

Peter and Pat
Albuquerque, NM, USA

Vic Soghomonian
northridge, CA, USA

Neil
Maple Valley,

Tim
Hamburg, Germany

Anonymous

Tania
Ottawa, Ontario,

Amy

FIELD PARTNER Learn more

Apoyo Integral administers this loan.

Social Performance Badges:

Family and Community Empowerment

Entrepreneurial Support

Facilitation of Savings

Innovation

What to know about this partner:

Field Partner:	Apoyo Integral
Field Partner Risk Rating	★★★★☆
Time on Kiva:	97 months
Kiva Borrowers:	18278
Total Loans:	$10,742,575
Interest & Fees are Charged	Yes
Average Cost to Borrower:	26% PY
Profitability (Return on Assets):	0.7%
Average Loan Size (% of Per Capita Income):	53.80%
Delinquency Rate:	4.89%
Loans at Risk Rate:	16.86%
Default Rate:	2.13%
Currency Exchange Loss Rate:	0.00%

More on this Field Partner »

MORE LOANS FROM THIS PARTNER
See All

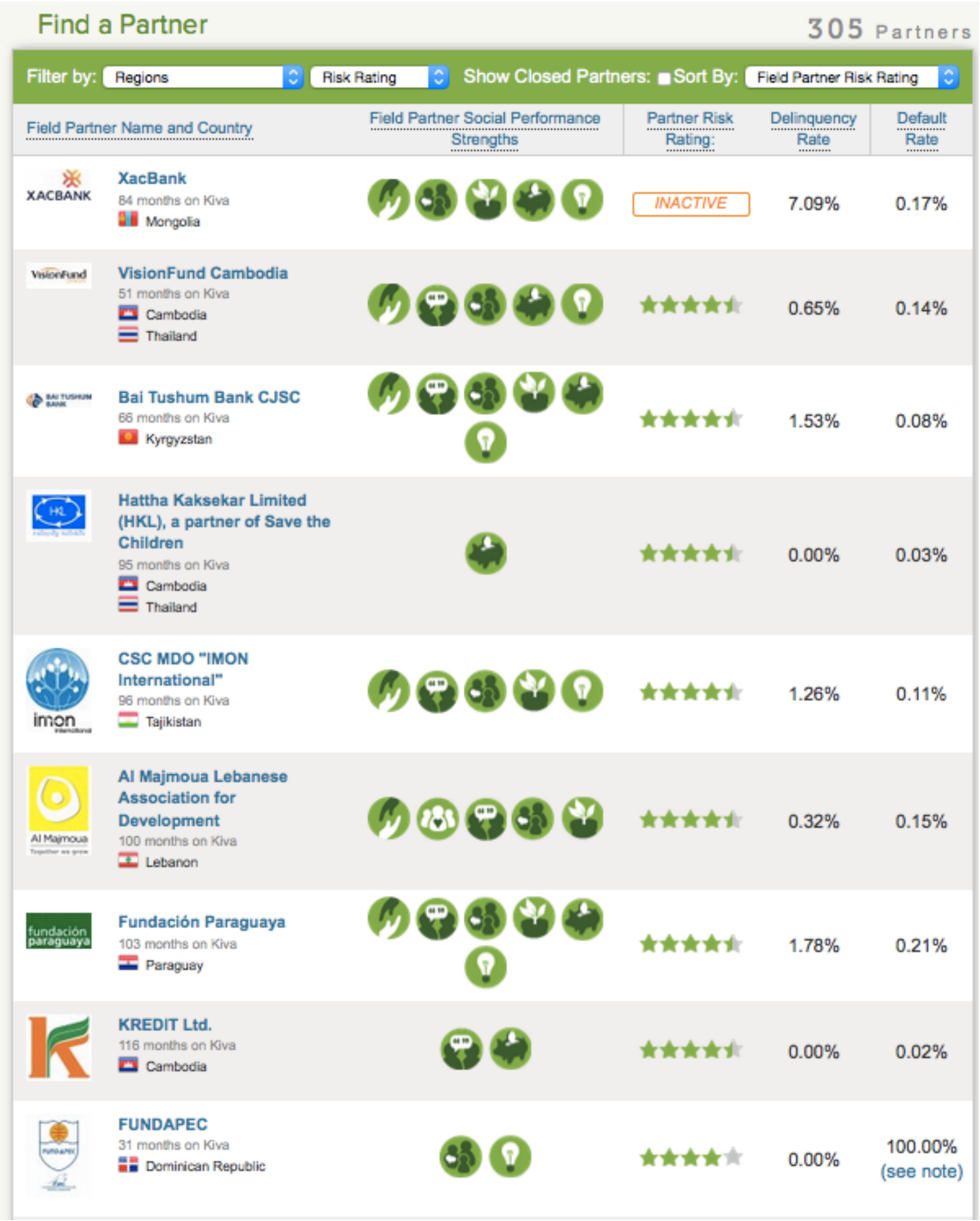
Fatima Guadalupe
Clothing
El Salvador
Raising funds
39% funded

Catalina Del Ro...
Farming
El Salvador
Raising funds
0% funded

Anonymous
Furniture Making
El Salvador
Raising funds
52% funded

Figure 2: Partial Kiva's Field Partners' list and Examples of Social Performance Accreditation Badges

The webpage URL is http://www.kiva.org/partners, visited on date January, 1st 2016.



Find a Partner

305 Partners

Filter by: Regions | Risk Rating | Show Closed Partners: □ Sort By: Field Partner Risk Rating

Field Partner Name and Country	Field Partner Social Performance Strengths	Partner Risk Rating:	Delinquency Rate	Default Rate
XacBank 84 months on Kiva Mongolia	●●●●●	INACTIVE	7.09%	0.17%
VisionFund Cambodia 51 months on Kiva Cambodia Thailand	●●●●●	★★★★½	0.65%	0.14%
Bai Tushum Bank CJSC 66 months on Kiva Kyrgyzstan	●●●●●●	★★★★½	1.53%	0.08%
Hattha Kaksekar Limited (HKL), a partner of Save the Children 95 months on Kiva Cambodia Thailand	●	★★★★½	0.00%	0.03%
CSC MDO "IMON International" 96 months on Kiva Tajikistan	●●●●●	★★★★½	1.26%	0.11%
Al Majmoua Lebanese Association for Development 100 months on Kiva Lebanon	●●●●●	★★★★½	0.32%	0.15%
Fundación Paraguaya 103 months on Kiva Paraguay	●●●●●●	★★★★½	1.78%	0.21%
KREDIT Ltd. 116 months on Kiva Cambodia	●●	★★★★½	0.00%	0.02%
FUNDAPEC 31 months on Kiva Dominican Republic	●●	★★★★☆	0.00%	100.00% (see note)

Figure 3: Social Performance Badges and Observed Loan Success

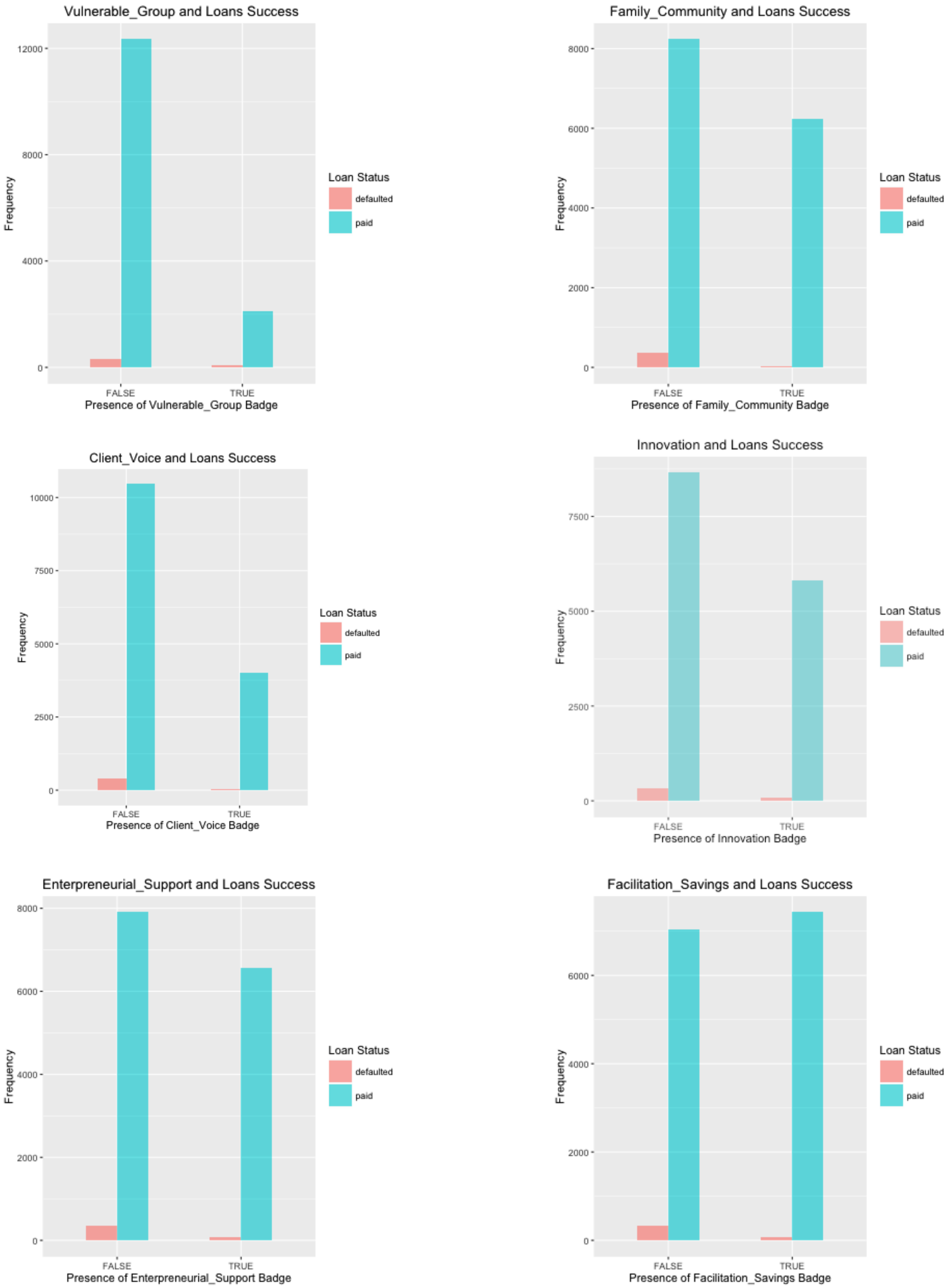


Figure 3B: Cross-tables of Loan Frequencies and Presence of Social Performance Badges

Antipoverty Badge	Repaid	Defaulted
FALSE	317	6966
TRUE	97	7597

Facilitation of Savings Badge	Repaid	Defaulted
FALSE	338	7030
TRUE	76	7433

Innovation Badge	Repaid	Defaulted
FALSE	333	8656
TRUE	81	5807

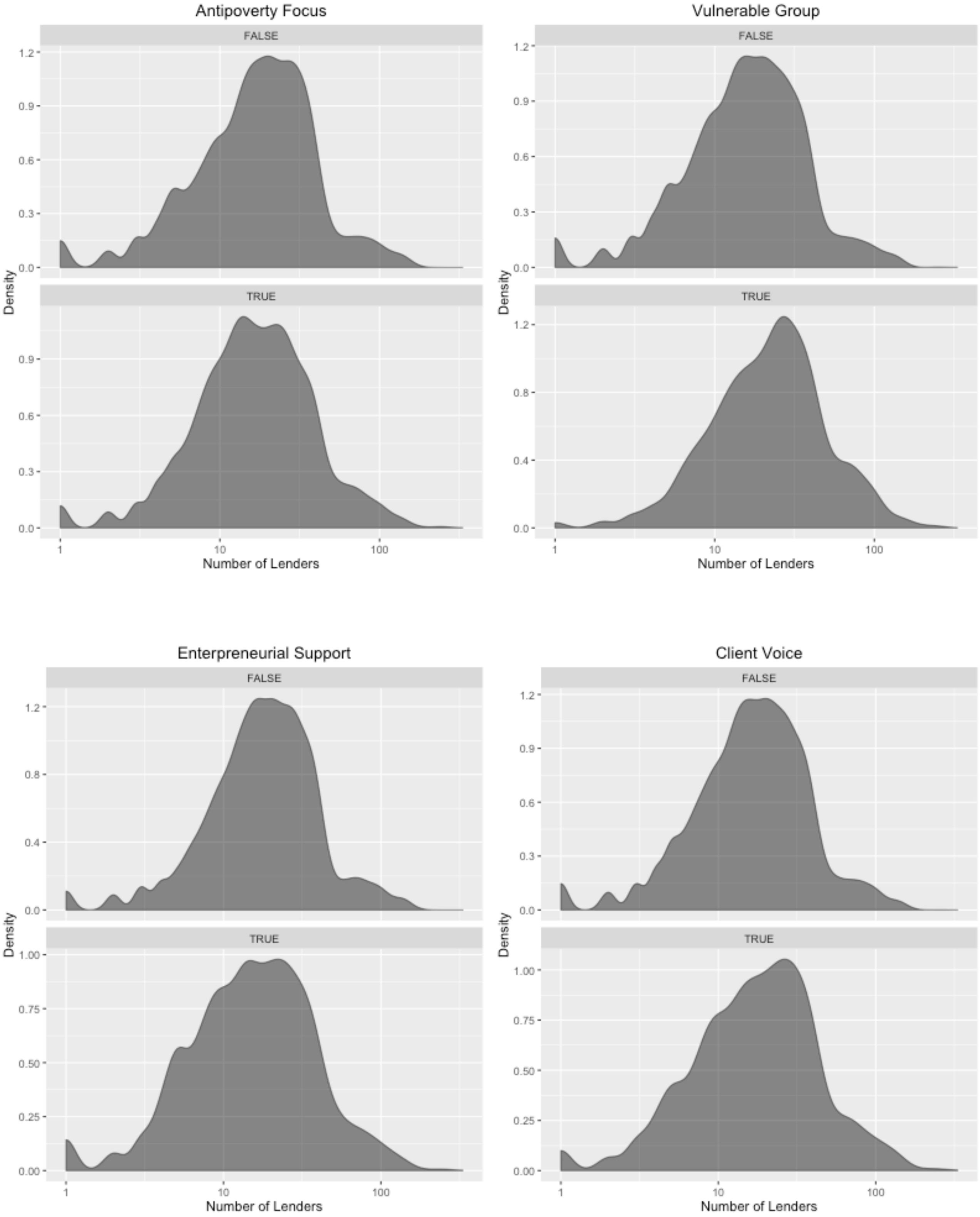
Client Voice Badge	Repaid	Defaulted
FALSE	383	10463
TRUE	31	4000

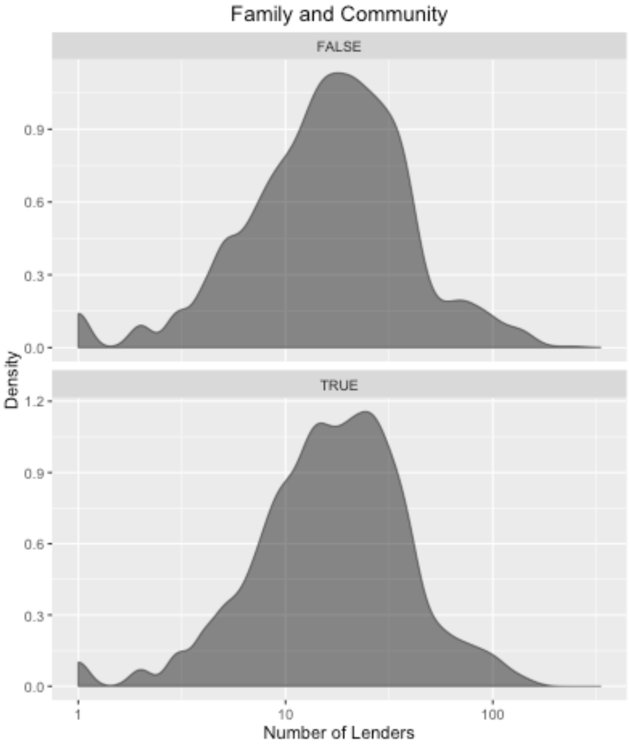
Family and Community Badge	Repaid	Defaulted
FALSE	379	8237
TRUE	35	6226

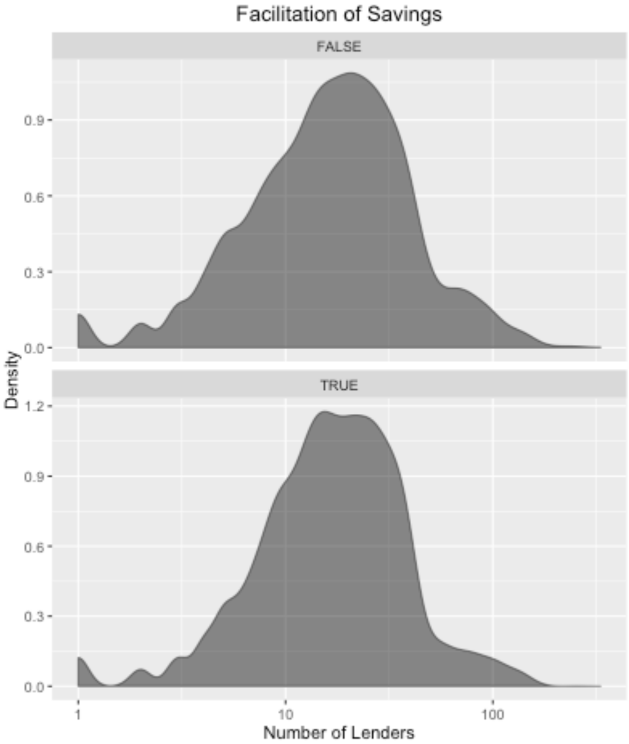
Vulnerable Group Badge	Repaid	Defaulted
FALSE	334	12346
TRUE	80	2117

Enterpreneurial Support Badge	Repaid	Defaulted
FALSE	345	7909
TRUE	69	6554

Figure 4: Density of Loans by Number of Lenders and Presence of Social Performance Badges




Family and Community
FALSE
TRUE
Density
0.0
0.3
0.6
0.9
1.2
1
10
100
Number of Lenders


Facilitation of Savings
FALSE
TRUE
Density
0.0
0.3
0.6
0.9
1.2
1
10
100
Number of Lenders

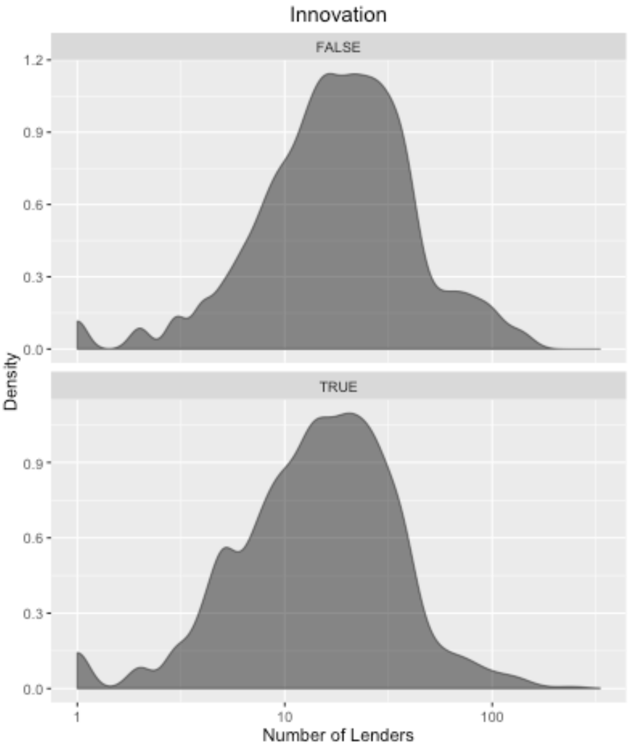
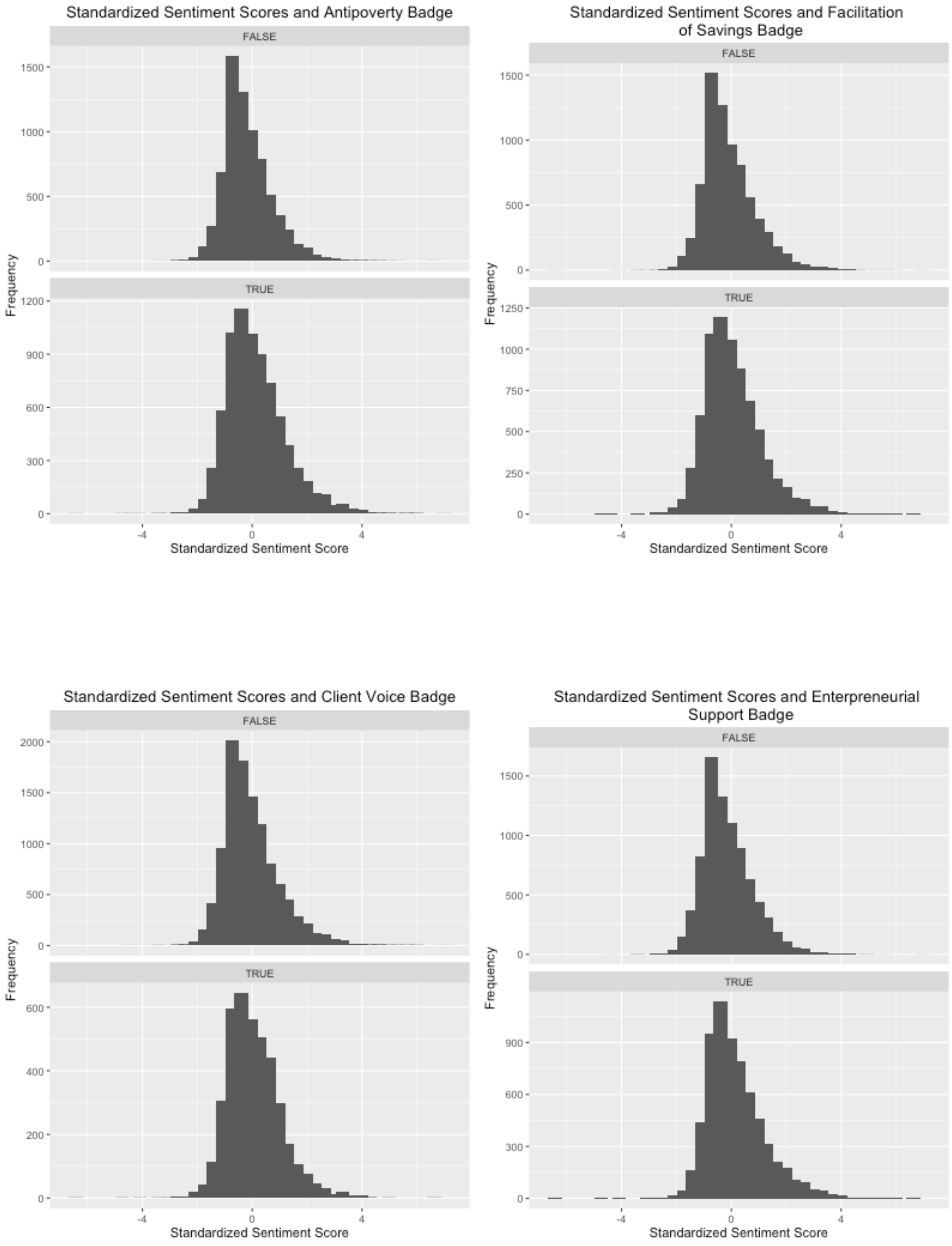
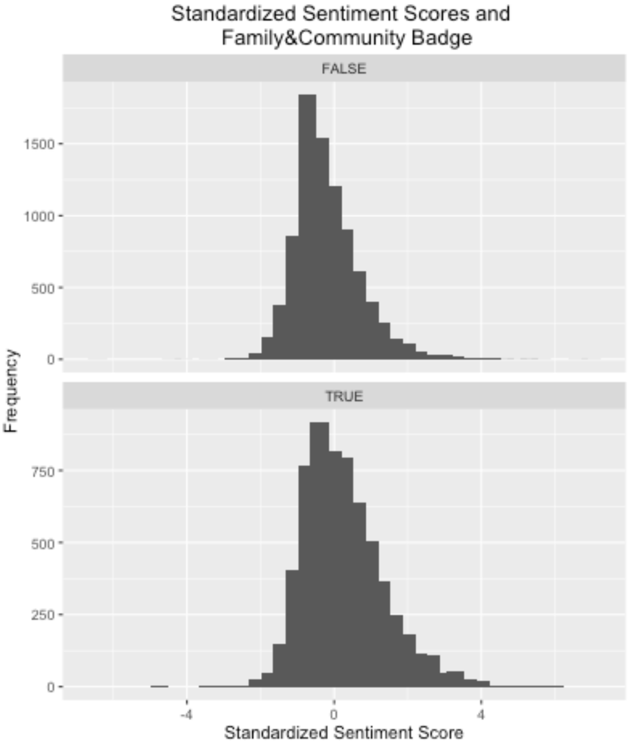
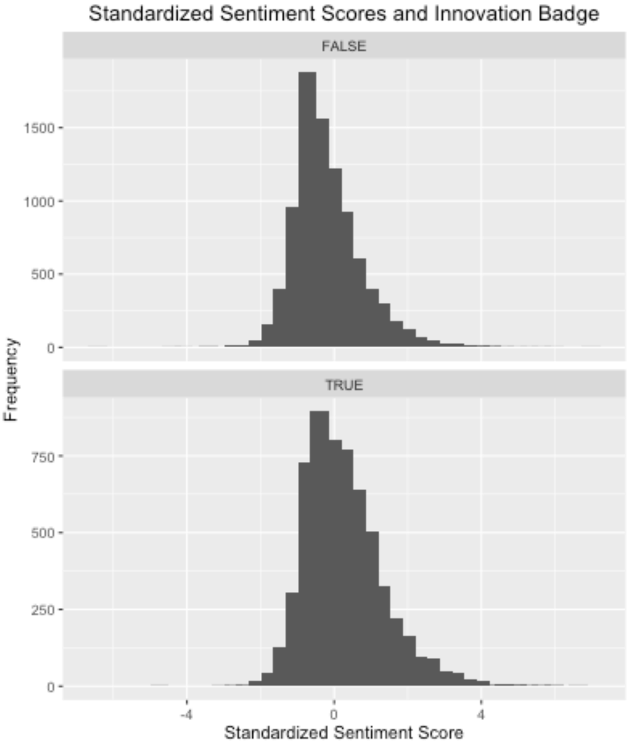

Innovation
FALSE
TRUE
Density
0.0
0.3
0.6
0.9
1.2
1
10
100
Number of Lenders

Figure 5: Sentiment Scores and Social Performance Badges - Distributions




Standardized Sentiment Scores and Family&Community Badge
FALSE
TRUE
Frequency
1500
1000
500
0
750
500
250
0
-4
0
4
Standardized Sentiment Score


Standardized Sentiment Scores and Innovation Badge
FALSE
TRUE
Frequency
1500
1000
500
0
750
500
250
0
-4
0
4
Standardized Sentiment Score

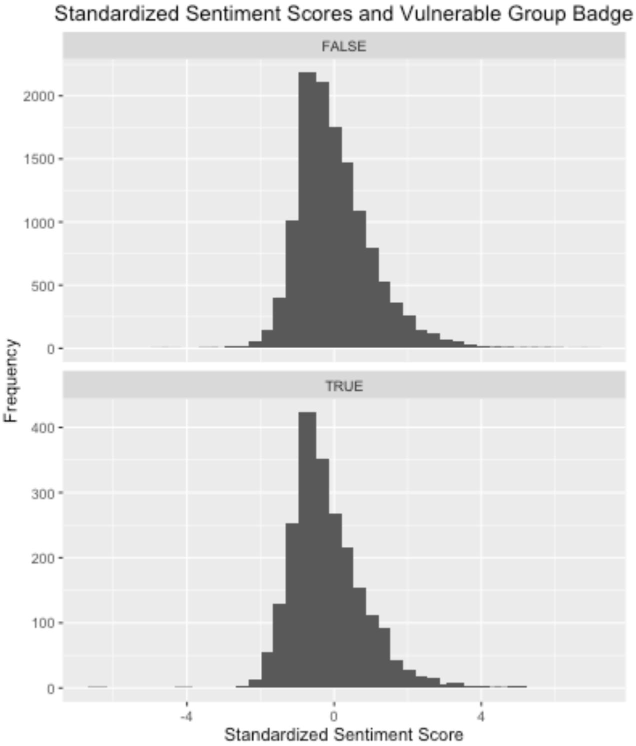
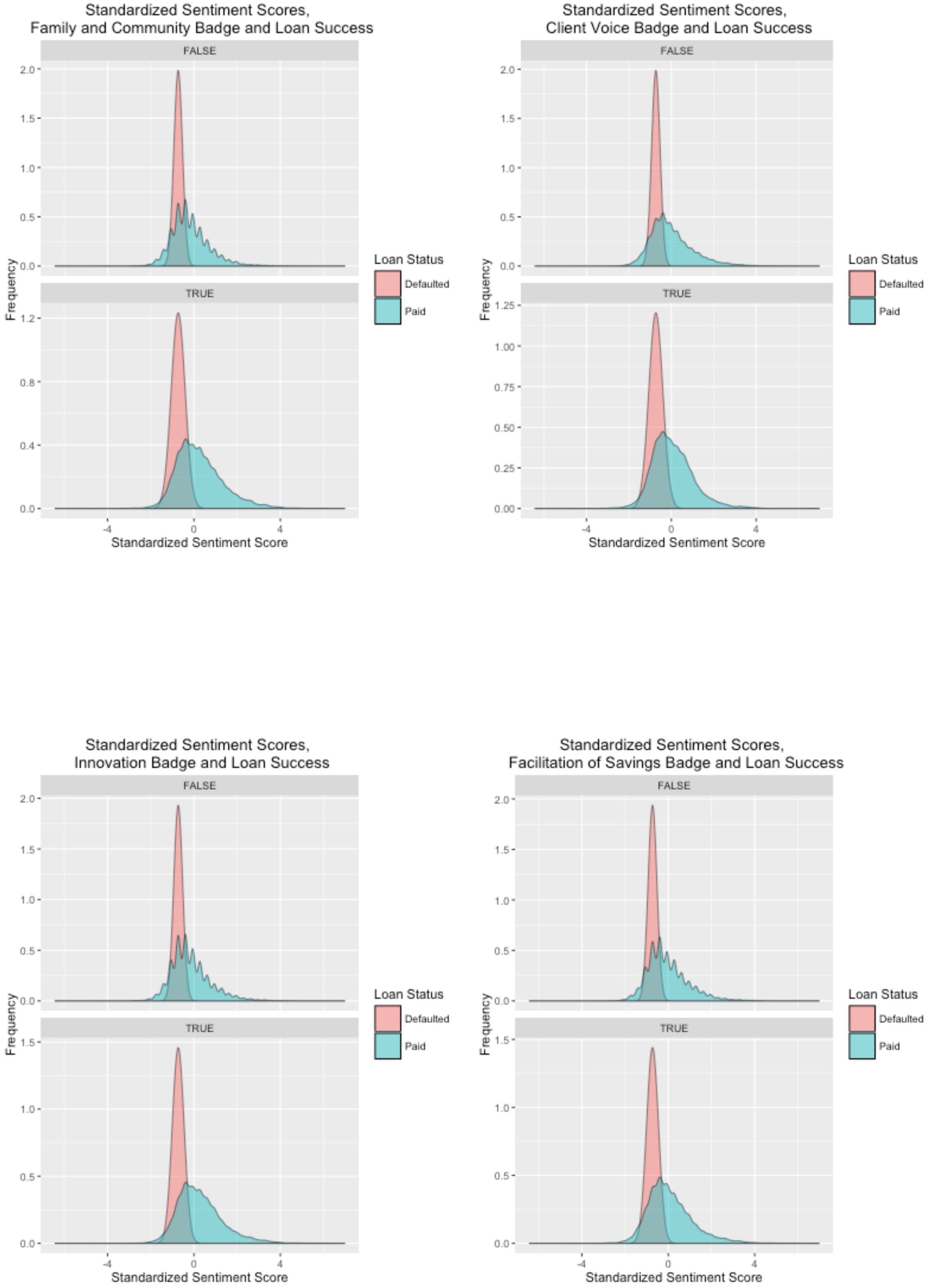
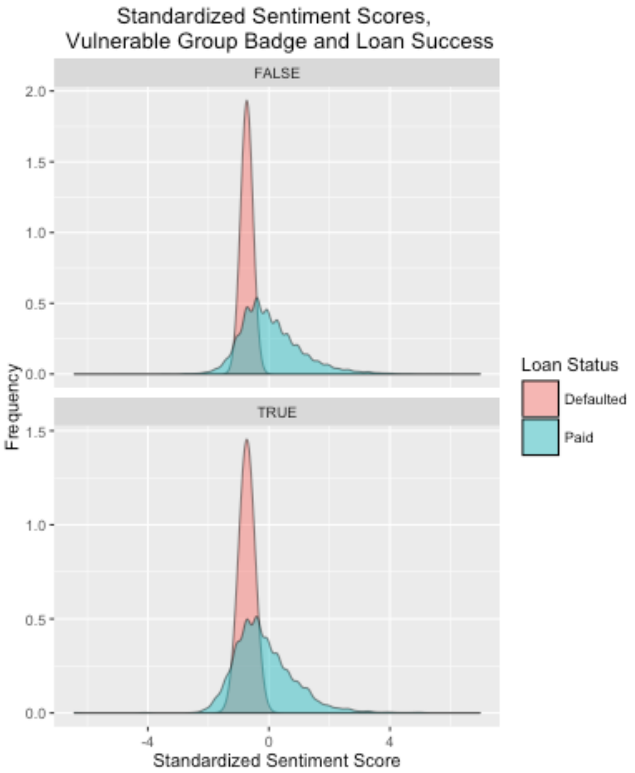
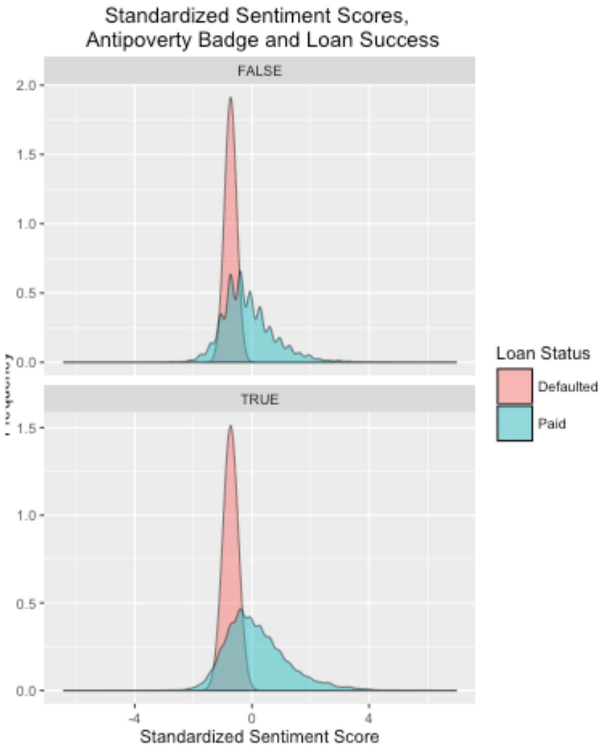

Standardized Sentiment Scores and Vulnerable Group Badge
FALSE
TRUE
Frequency
2000
1500
1000
500
0
400
300
200
100
0
-4
0
4
Standardized Sentiment Score

Figure 6: Sentiment Score, Loan Success and Social Performance Badges




Standardized Sentiment Scores, Vulnerable Group Badge and Loan Success


Standardized Sentiment Scores, Antipoverty Badge and Loan Success

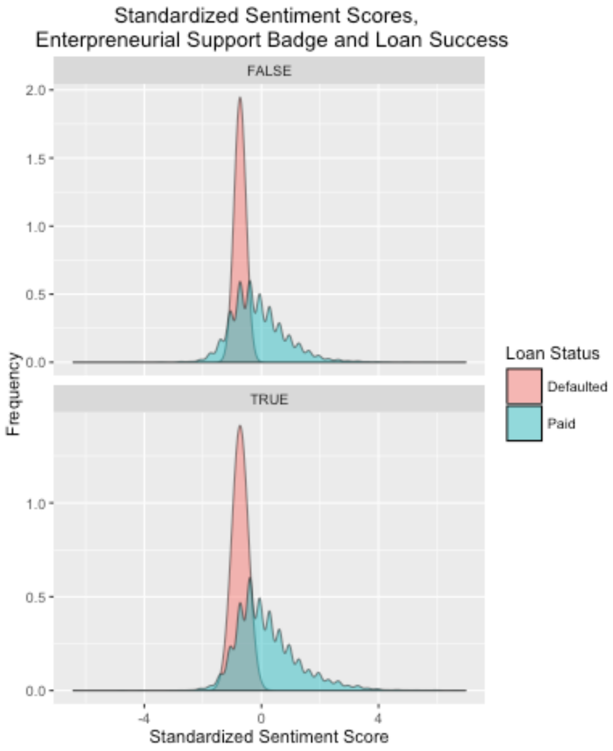

Standardized Sentiment Scores, Enterpreneurial Support Badge and Loan Success

Figure 7: Optimal Prediction Threshold - Model 0

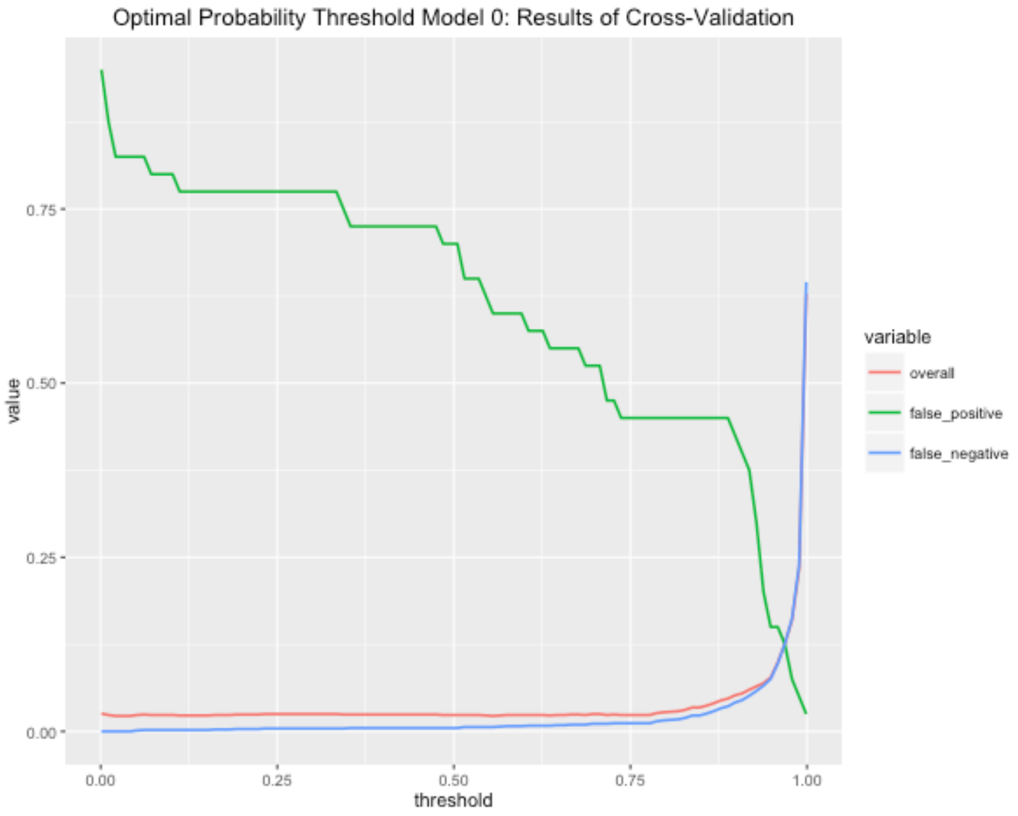


Figure 8A: Log Lambda and Mean Squared Error - Model 1 - Lambda Free to Vary

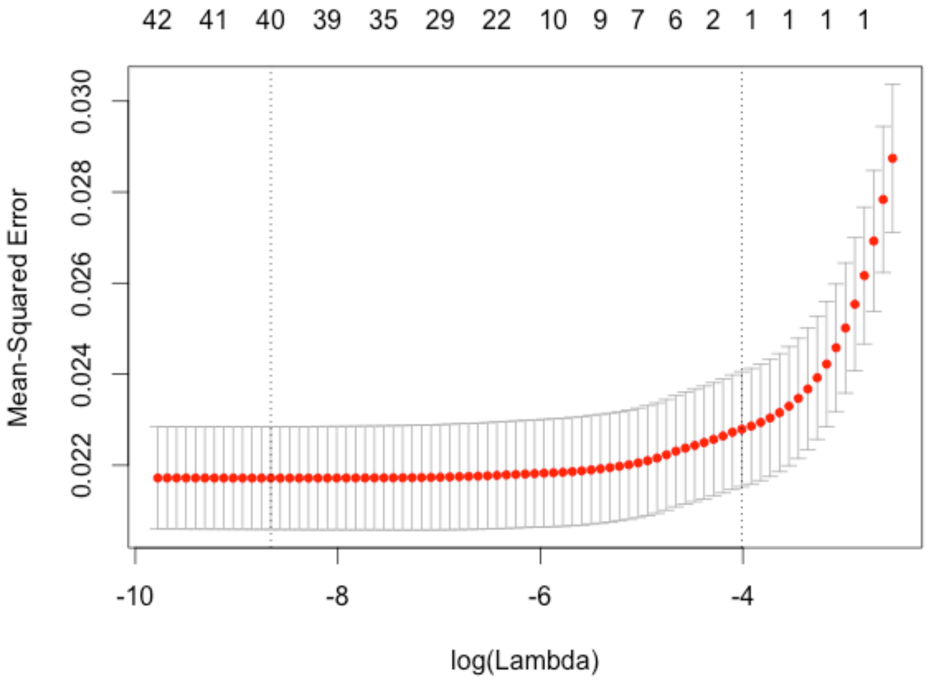


Figure 8B: Log Lambda and Mean Squared Error - Model 1 - Grid of Lambda Values

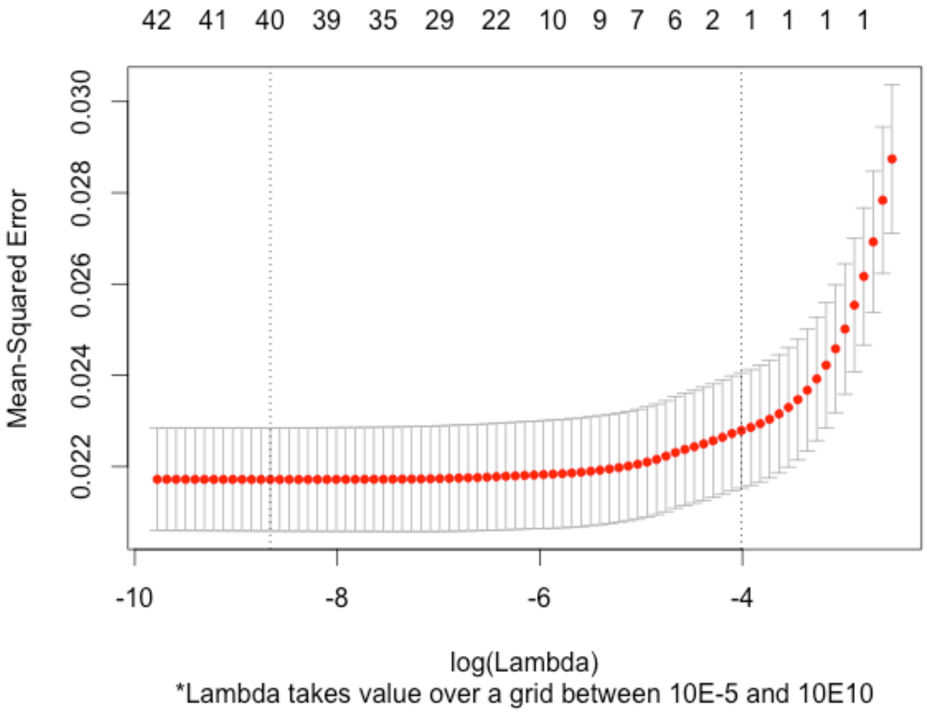


Figure 9: Optimal Prediction Threshold - Model 1

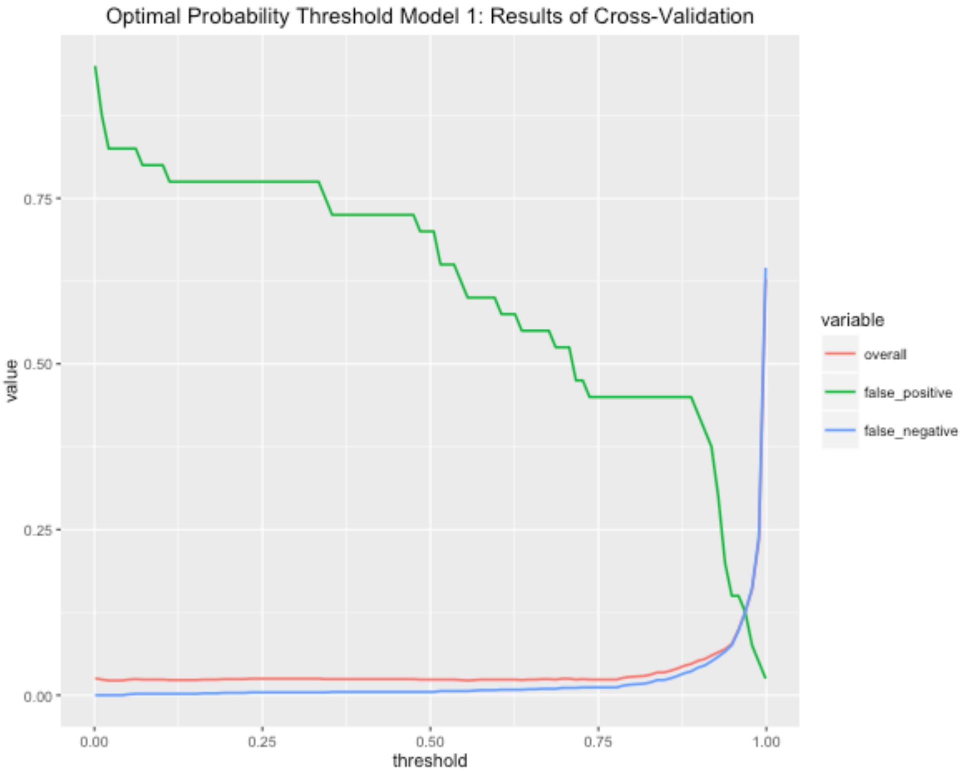


Figure 10: Optimal Prediction Threshold - Model A

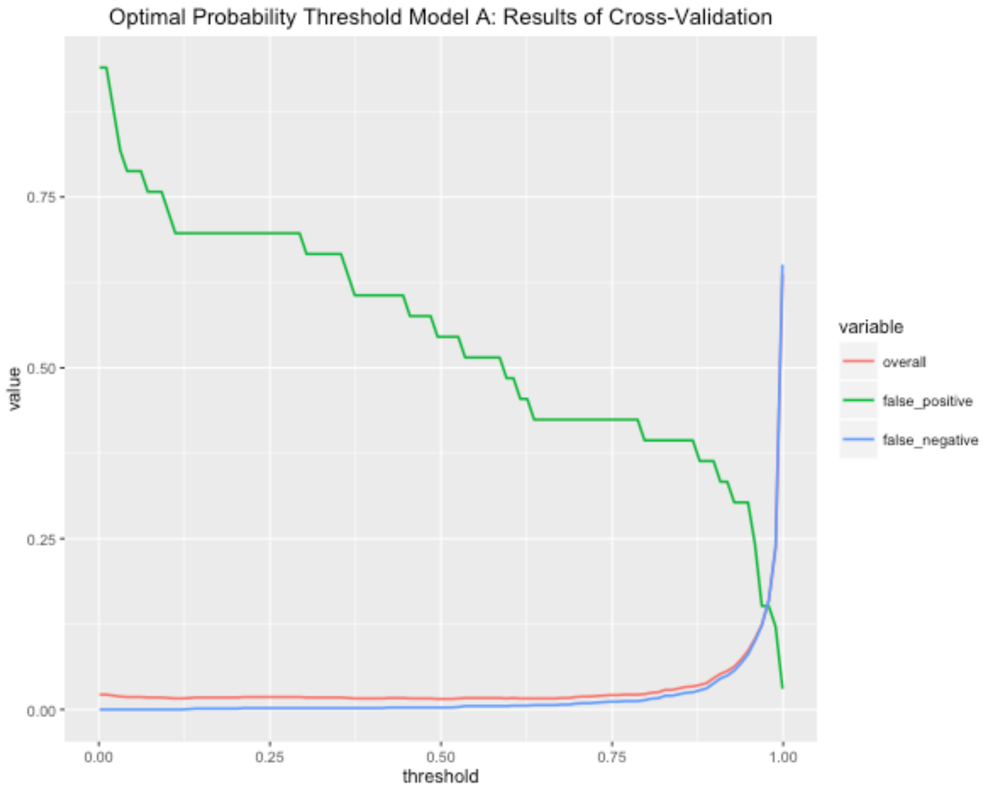


Figure 11: Parameter Shrinkage Plot - Model B

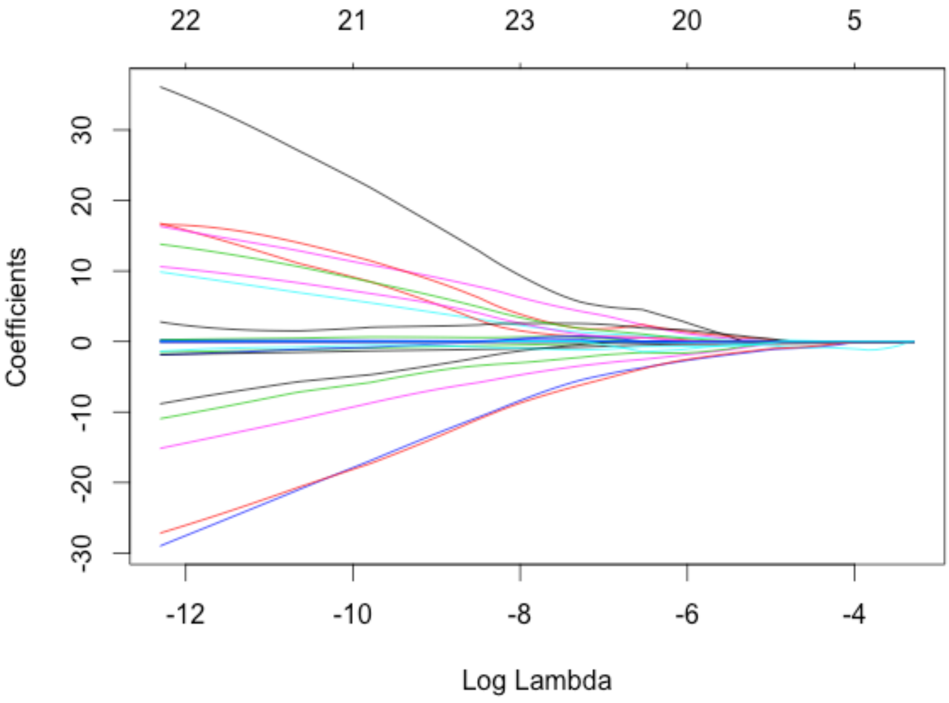


Figure 12: Log Lambda and Mean Squared Error - Model B - Lambda Free to Vary

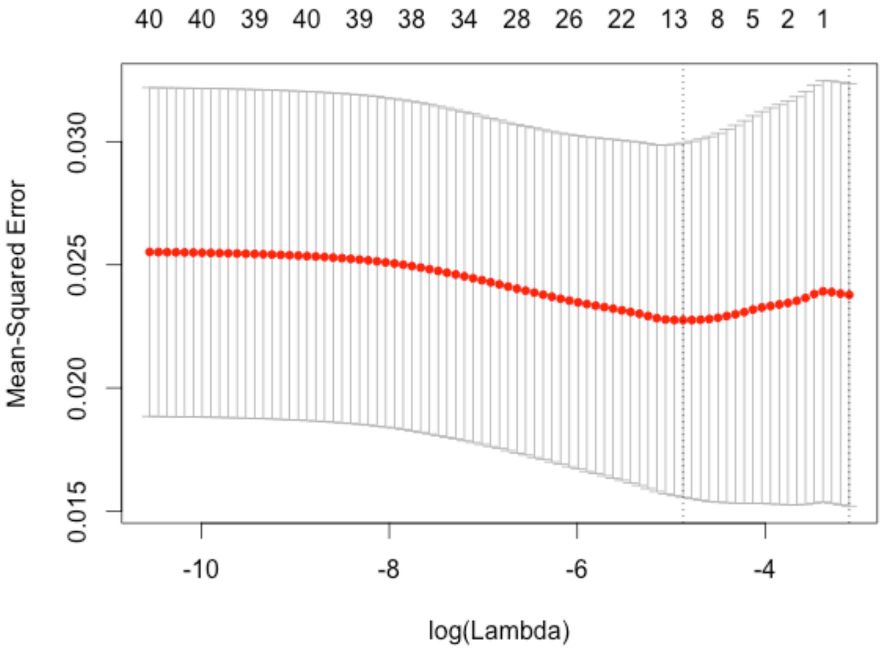


Figure 13: Optimal Prediction Threshold - Model C

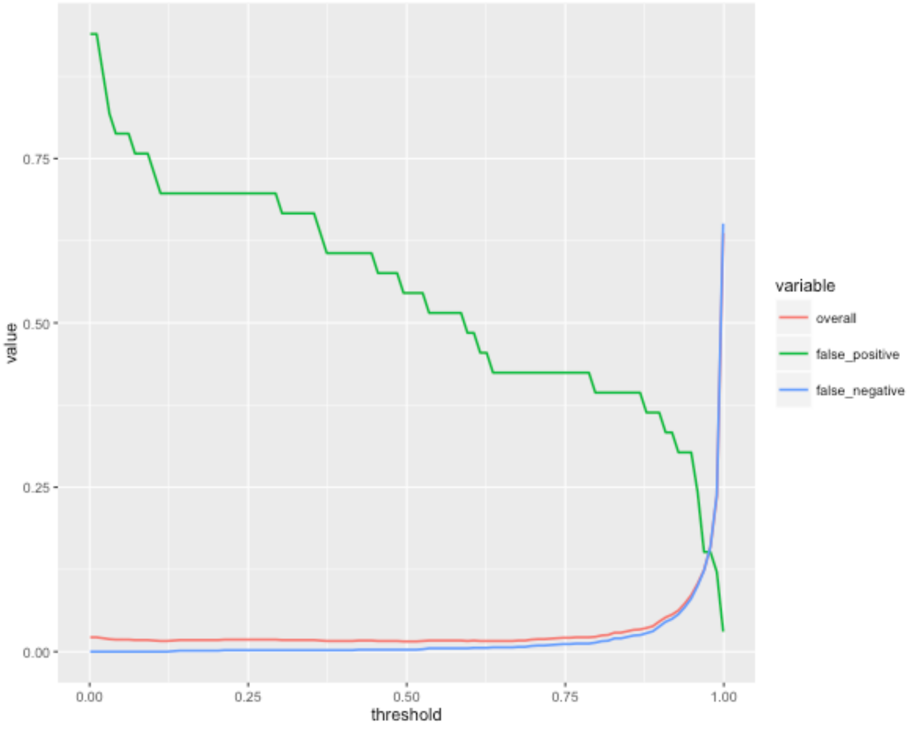


Table 3: Ouput from Model A

Coefficients	**Estimate**	**Std. Error**	**z value**	**A.E.O.**
(Intercept)	3,733	0,787	4,745	41,817
loans.sectorArts	-0,100	0,594	-0,168	0,905
loans.sectorClothing	-0,924	0,282	-3,274	0,397
loans.sectorConstruction	-0,050	0,557	-0,089	0,952
loans.sectorEducation	-0,009	0,721	-0,013	0,991
loans.sectorEntertainment	-1,937	0,862	-2,248	0,144
loans.sectorFood	-0,528	0,225	-2,351	0,590
loans.sectorHealth	-0,566	0,608	-0,929	0,568
loans.sectorHousing	0,380	0,501	0,759	1,462
loans.sectorManufacturing	0,308	0,838	0,367	1,360
loans.sectorPersonal Use	0,131	1,093	0,120	1,141
loans.sectorRetail	-0,651	0,237	-2,745	0,522
loans.sectorServices	-0,237	0,313	-0,758	0,789
loans.sectorTransportation	0,286	0,511	0,560	1,331
loans.sectorWholesale	-1,107	0,856	-1,292	0,331
loans.loan_amount	0,002	0,000	3,444	1,002
loans.lender_count	-0,049	0,014	-3,470	0,952
loans.bonus_credit_eligibilityTRUE	-1,026	0,298	-3,443	0,359
loans.terms.repayment_intervalIrregularly	0,839	0,532	1,577	2,313
loans.terms.repayment_intervalMonthly	0,821	0,415	1,980	2,274
loans.terms.repayment_term	-0,124	0,015	-8,110	0,884
loans.terms.loss_liability.nonpaymentpartner	1,882	0,181	10,394	6,566
loans.journal_totals.entries	-0,435	0,066	-6,587	0,647
partners.statusclosed	0,902	0,638	1,414	2,464
partners.statusinactive	14,077	376,540	0,037	1E+06
partners.statuspaused	-0,140	0,831	-0,169	0,869
partners.rating	0,975	0,230	4,232	2,651
partners.delinquency_rate	-0,022	0,012	-1,864	0,978
partners.default_rate	-0,194	0,011	-18,276	0,823
partners.total_amount_raised	0,000	0,000	0,340	1,000
partners.loans_posted	0,000	0,000	1,909	1,000
partners.average_loan_size_percent_per_capita_income	-0,010	0,001	-7,106	0,991
partners.loans_at_risk_rate	0,026	0,010	2,646	1,027
partners.portfolio_yield	-0,003	0,007	-0,431	0,997
partners.profitability	-0,004	0,012	-0,347	0,996
score	0,470	0,035	13,442	1,599

A.E.O.: Average Effect on the Odds of success = e^β

Chi-square: 1753.038

Df: 35

Associated p-value: 0.000

Table 4: Output from Model B

Coefficients	Estimate
(Intercept)	**5,20086**
loans.sectorArts	0,00000
loans.sectorClothing	0,00000
loans.sectorConstruction	0,00000
loans.sectorEducation	0,00000
loans.sectorEntertainment	0,00000
loans.sectorFood	0,00000
loans.sectorHealth	0,00000
loans.sectorHousing	0,00000
loans.sectorManufacturing	**-0,82509**
loans.sectorPersonal.Use	0,00000
loans.sectorRetail	**-0,15297**
loans.sectorServices	0,00000
loans.sectorTransportation	0,00000
loans.sectorWholesale	0,00000
partners.statusclosed	0,00000
partners.statusinactive	0,00000
partners.statuspaused	0,00000
partners.charges_fees_and_interestTRUE	0,00000
recent.loans.loan_amount	0,00000
recent.loans.lender_count	**-0,00005**
recent.loans.terms.repayment_term	**-0,00325**
recent.loans.journal_totals.entries	**-0,66475**
recent.partners.rating	0,00000
recent.partners.delinquency_rate	**-0,01626**
recent.partners.default_rate	**-0,29403**
recent.partners.total_amount_raised	0,00000
recent.partners.loans_posted	0,00000
recent.partners.average_loan_size_percent_per_capita_income	**-0,00390**
recent.partners.loans_at_risk_rate	0,00000
recent.partners.portfolio_yield	**-0,00303**
recent.partners.profitability	**-0,00538**
recent.score	**0,12955**
loans.bonus_credit_eligibilityTRUE	0,00000
loans.terms.repayment_intervalIrregularly	0,00000
loans.terms.repayment_intervalMonthly	0,00000
loans.terms.loss_liability.nonpaymentpartner	0,00000
AntipovertyTRUE	0,00000
Vulnerable_GroupTRUE	0,00000
Client_VoiceTRUE	**-0,00603**
Family_CommunityTRUE	0,00000
Enterpreneurial_SupportTRUE	0,00000
Facilitation_SavingsTRUE	0,00000
Innovation_TRUE	0,00000

Table 5: Comparison Between Model 1, Model A and Model B

Coefficients	Estimates Model 1	Estimates Model A	Estimates Model B
(Intercept)	5,220	3,733	5,201
loans.loan_amount	0,001	0,002	0,000
loans.lender_count	-0,043	-0,049	-0,00005
loans.bonus_credit_eligibilityTRUE	-0,873	-1,026	0,000
loans.terms.repayment_intervalIrregularly	1,292	0,839	0,000
loans.terms.repayment_intervalMonthly	1,165	0,821	0,000
loans.terms.repayment_term	-0,134	-0,124	-0,003
loans.terms.loss_liability.nonpaymentpartner	1,823	1,882	0,000
loans.journal_totals.entries	-0,540	-0,435	-0,665
partners.rating	0,981	0,975	0,000
partners.delinquency_rate	-0,027	-0,022	-0,016
partners.default_rate	-0,204	-0,194	-0,294
partners.total_amount_raised	0,000	0,000	0,000
partners.loans_posted	0,000	0,000	0,000
partners.average_loan_size_percent_per_capita_income	-0,012	-0,010	-0,004
partners.loans_at_risk_rate	0,033	0,026	0,000
partners.portfolio_yield	-0,011	-0,003	-0,003
partners.profitability	**0,017**	**-0,004**	**-0,005**
partners.charges_fees_and_interest	0,000	NA	0,000
AntipovertyTRUE	-0,447	-	0,000
Vulnerable_GroupTRUE	-0,656	-	0,000
Client_VoiceTRUE	**0,169**	-	**-0,006**
Family_CommunityTRUE	1,213	-	0,000
Enterpreneurial_SupportTRUE	0,999	-	0,000
Facilitation_SavingsTRUE	0,139	-	0,000
InnovationTRUE	-0,649	-	0,000
score	**0,457**	**0,470**	**0,130**
dummy_Arts	0,000	-0,100	0,000
dummy_Food	-0,473	-0,528	0,000
dummy_Services	-0,356	-0,237	0,000
dummy_Agriculture	0,200	0,000	0,000
dummy_Retail	-0,706	-0,651	-0,153
dummy_Construction	0,146	-0,050	0,000
dummy_Clothing	-1,024	-0,924	0,000
dummy_Housing	0,543	0,380	0,000
dummy_Wholesale	-1,195	-1,107	0,000
dummy_Manufacturing	**-0,148**	**0,308**	**-0,825**
dummy_Health	-0,450	-0,566	0,000
dummy_Personal_Use	0,238	0,131	0,000
dummy_Entertainment	-1,949	-1,937	0,000
dummy_Education	**0,221**	**-0,009**	**0,000**
dummy_Transportation	0,000	0,286	0,000
dummy_active	-1,260	0,000	0,000
dummy_paused	-2,591	-0,140	0,000
dummy_inactive	12,013	14,077	0,000

APPENDIX 2: Predicting Loan Success Using Random Forests

1. Introduction: Why a Logistic Regression May Not Be Enough

The loan repayment classification problem discussed in the present research has been extensively addressed using simple logistic regression models and lasso regressions. However, there are other ways in which the research problem can be tackled. One of the alternatives is the use decision trees and tree ensembles. There are several reasons for which the use of these tools may be appropriate for the purposes of the present research. First of all, they are easily interpretable while mirroring human decision-making. This could be particularly relevant, given the high degree of personal itneractions and interconnectedness of the parties in the prosocial lending setting. Second, the potential issue of the high number of predictors in the loan classification problem is automatically addressed, because decision trees and ensembles perform variable selection implicitly. The analyses performed throughout the present research have always included a minimum of 27 individual predictors, up to more than 50 in the more complex models. This problem could be easily addressed by a tree ensemble. Eventually, these instruments would eliminate the need for too many comparative models and analyses, since interaction effects are automatically taken into account. For these reasons, this final section attempts to handle the loan repayment classification problem using a non-linear technique: a random forest. Random forests are preferred over individual trees and other bagging procedures because, by building decorrelated trees, the problem of model variance is greatly reduced.

2. Model Fit and Accuracy

All the variables considered for the Model 0 linear regression (Section 5) are used to fit the random forest. Two different models are then compared. The first model (named rf_5), uses the default number of variables randomly sampled as candidates at each split. For the present case, it would be the rounded square root of the total number of predictors (5). The second model (named rf_9) uses one third of the total number of predictors, which, for the present case, would be 9. A total of 1000 trees to grow is selected. **Figure 14** shows the out-of-bag error estimate for the fitted models and that the lowest error estimate is achieved by Model rf_9 (green line).

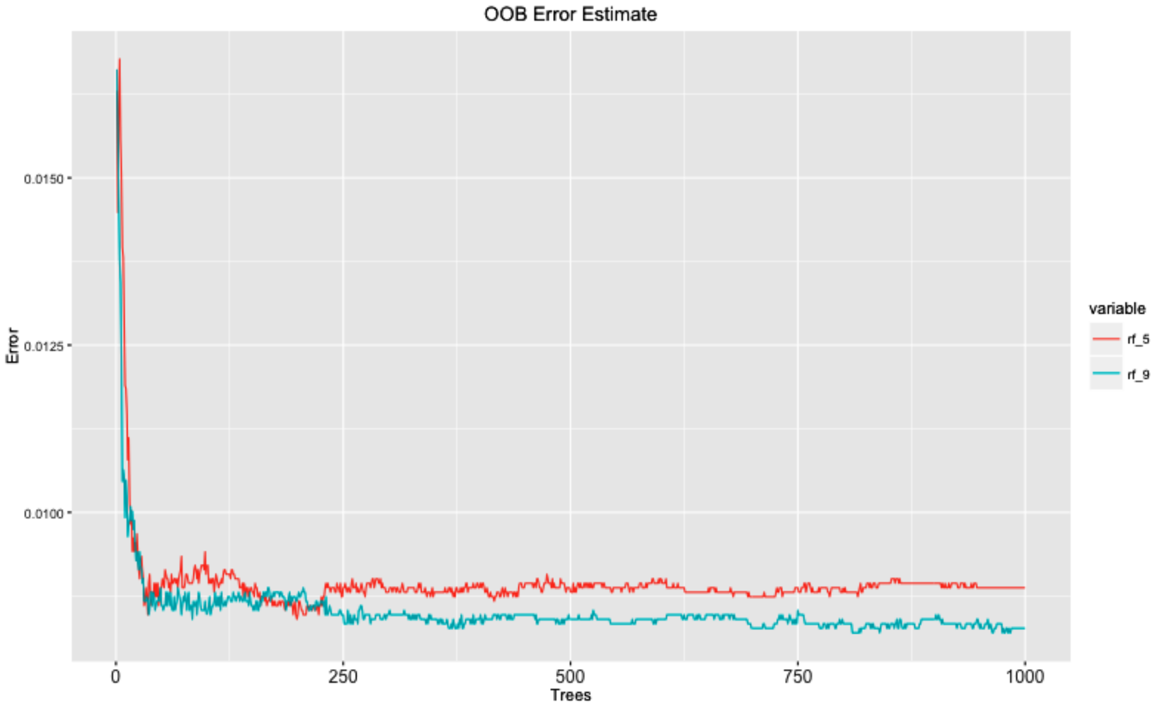


FIGURE 14: OOB ERROR ESTIMATES FOR RF_5 AND RF_9

Tables 6A and 6B, instead, show the prediction accuracy of Model rf_9 and the associated misclassification rate, assuming that all the errors have the same weight in the associated cost function.

Observed*Predicted Loan Status	Defaulted	Paid
Defaulted	318	96
Paid	27	14436

Random Forest - rf_9	
mtry = 9, n = 14877, ntree = 1000	
Prediction Accuracy	99,17%
Misclassification Rate	0,83%

TABLE 6A: CONFUSION MATRIX FOR RF_9. TABLE 6B: PREDICTION ACCURACY AND MISCLASSIFICATION RATE FOR RF_9

Based on the output shown in Tables 15A and 15B, the random forest achieves a better result,compared to the simple logistic and lasso regressions in Sections 5, 6 and 7. The prediction accuracy is further increased, reaching the point of 99,17%. The misclassification rate is below 1%, leading to the conclusion that the model has satisfactory predictive power.

3. Variable Importance

An interesting output from the random forest is the graph of variable importance. **Figure 15** shows the variables used in the model, sorted in descending order of importance to the procedure.

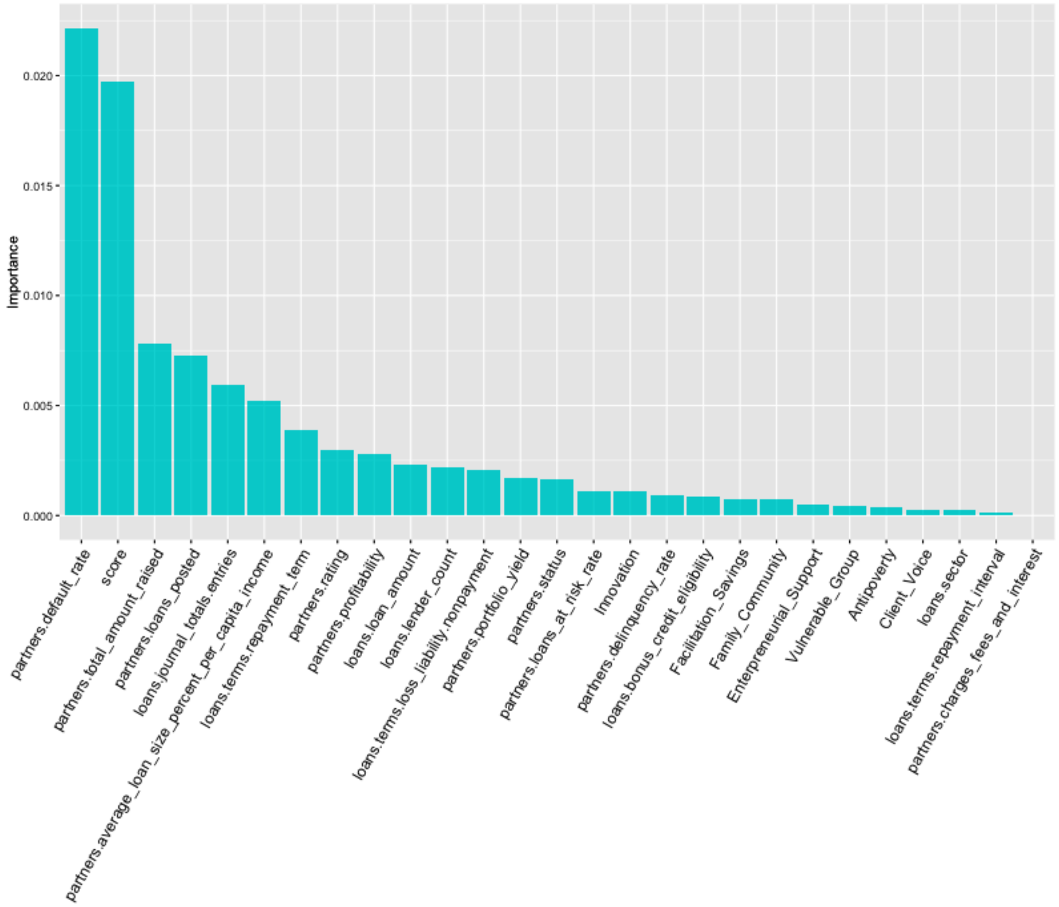


FIGURE 15: VARIABLE IMPORTANCE FOR THE RF_9 PROCEDURE

Interestingly, from Figure 15 it is possible to notice that the sentiment score of text descriptions is the second most important variable in the procedure. The first most important variable is the default rate of the Field Partner. Among the Social Performance Badges, the most important is the Innovation badge, which, however, scores low relative to the other loan- and partner-specific predictors. **Table 7** reports the variable importance values in detail.

Variable	Importance
partners.default_rate	0,0222
score	0,0197
partners.total_amount_raised	0,0078
partners.loans_posted	0,0073
loans.journal_totals.entries	0,0059
partners.average_loan_size_percent_per_capita_income	0,0052
loans.terms.repayment_term	0,0039
partners.rating	0,0030
partners.profitability	0,0028
loans.loan_amount	0,0023
loans.lender_count	0,0022
loans.terms.loss_liability.nonpayment	0,0021
partners.portfolio_yield	0,0017
partners.status	0,0016

Variable	Importance
partners.loans_at_risk_rate	0,0011
Innovation	0,0011
partners.delinquency_rate	0,0009
loans.bonus_credit_eligibility	0,0009
Facilitation_Savings	0,0008
Family_Community	0,0007
Enterpreneurial_Support	0,0005
Vulnerable_Group	0,0005
Antipoverty	0,0004
Client_Voice	0,0003
loans.sector	0,0002
loans.terms.repayment_interval	0,0001
partners.charges_fees_and_interest	0,0000

TABLE 7: VARIABLE IMPORTANCE - TABLE OUTPUT

From Table 7, it appears that the variable "partners.charges_fees_and_interest" has no importance in the classification procedure, which confirms the findings from the lasso regression performed in Section 6 of the present research. The variable "loans_sector" is also relatively unimportant: this is reinforced by the fact that, in Section 6, the lasso regression penalized and set to zero the coefficients associated with two levels of the variable. On the other hand, almost all the Social Performance Badge variables appear to have little importance in the procedure, which somewhat contradicts the findings from Sections 5 and 6 of the present research. Finally, as already highlighted by the bar plot in Figure 15, the sentiment score of text descriptions is once again confirmed as an important predictor of loan successful repayment.

4. Model Relevance and Conclusions for Practitioners

The random forest procedure performed on the same set of data as Model 0 improved the loan repayment prediction accuracy by more than 1%, with a reduction in the number of false positives of 17 units. The implication for practitioners may be that the random forest could more suitable and appropriate for purely predictive purposes. The procedure, in fact, is more accurate, easy to interpret and accounts automatically for possible interaction effects. On the other hand, logistic and lasso regressions could be more appropriate to complement and expand the present work with further research. The two linear analytical methods, in fact, allowed for a more detailed interpretation of the impact of individual variables on the chances of loan repayment, while achieving a satisfactory level of prediction accuracy (97,97%). This can be particularly useful, for example, to test for the effect of accreditation badges in contexts other than Kiva.